Fourth Baltic-Nordic Conference on Survey Statistics

PROCEEDINGS

24–28 August 2015, Helsinki, Finland

UNIVERSITY OF HELSINKI

Statistics Finland

FINNISH STATISTICAL SOCIETY

4th Baltic-Nordic Conference on Survey Statistics
24–28 August 2015, Helsinki, Finland

Program Committee

Juha Alho (University of Helsinki)
Natallia Bokun (Belarus State Economic University)
Michael Carlson (Stockholm University)
Kari Djerf (Statistics Finland)
Tetiana Ianevich (Taras Shevchenko National University of Kyiv)
Danutė Krapavickaitė (Vilnius Gediminas Technical University)
Gunnar Kulldorff (Umeå University, Honorary Member, until 25 June 2015)
Janis Lapins (Bank of Latvia)
Anna Larchenko (Ministry of Foreign Affairs of the Republic of Belarus)
Risto Lehtonen (University of Helsinki, Chair)
Mārtiņš Liberts (Statistics Latvia)
Aleksandras Plikusas (Vilnius University)
Kaja Sõstra (Statistics Estonia)
Daniel Thorburn (Stockholm University)
Imbi Traat (University of Tartu)
Olga Vasylyk (Taras Shevchenko National University of Kyiv)

Organizing Committee

Risto Lehtonen (University of Helsinki, chair)
Faiz Alsuhail (Statistics Finland)
Jaana Lehtinen (University of Helsinki)
Jyrki Möttönen (Finnish Statistical Society)
Maria Valaste (The Social Insurance Institution of Finland)
Kimmo Vehkalahti (University of Helsinki)
Jenna Wallenius (University of Helsinki)

Sponsors

International Association of Survey Statisticians (IASS)
Nordic Council of Ministers (Nordplus HE Programme)
Federation of Finnish Learned Societies (TSV)
Statistics Finland
University of Helsinki
SAS Institute

ISBN 978-951-51-1469-3 (Printed version)
ISBN 978-951-51-1470-9 (PDF version)

Photo of the cover: Niko Soveri / Visit Helsinki

Foreword

This proceedings publication includes the abstracts of papers of the Fourth Baltic-Nordic Conference on Survey Statistics, BaNoCoSS 2015, taking place on 24-28 August 2015 in Helsinki, Finland. The BaNoCoSS series of events constitutes of international scientific conferences presenting recent developments of the theory, methodology and application of survey statistics in a wide sense. Previous conferences were organized in 2002 in Ammarnäs, Sweden, 2007 in Kuusamo, Finland and 2011 in Höga Kusten, Sweden.

The conference is organized by the Baltic-Nordic-Ukrainian Network on Survey Statistics together with the University of Helsinki, Statistics Finland and Finnish Statistical Society. Since 1997, the Baltic-Nordic-Ukrainian Network on Survey Statistics has arranged a yearly event − Workshop, Summer School or BaNoCoSS conference − together with partner organizations in one of the network countries, Belarus, Estonia, Finland, Latvia, Lithuania, Sweden or Ukraine. The activity was initiated, expanded and guided for several years by Professor Gunnar Kulldorff, University of Umeå. I regret to inform that he passed away this summer. We miss the presence and contribution of this extraordinary person.

The present conference appears to be one of the most extensive of the network events this far. There are over 90 registered participants coming from 14 different countries. This year, the share of Finnish participants is large showing increasing interest in survey statistics in Finland. The program includes close to 70 invited and contributed papers, covering broadly the topics in modern survey statistics and official statistics. The educational flavour of the conference is strong: we have a privilege to follow the keynote lectures given by three prominent statisticians, Professor Jelke Bethlehem of University of Leiden, Professor Partha Lahiri of University of Maryland and Dr Eric Schulte-Nordholt, Senior Researcher and Project Leader at Statistics Netherlands. We are glad to have several additional invited speakers, including Yves Berger of University of Southampton, Yan Li of University of Maryland, Kaija Ruotsalainen of Statistics Finland and Li-Chun Zhang of University of Southampton and Statistics Norway, just to mention a few of them.

The conference would not have realized without the contribution of several people and organizations. I want to express thanks to the members of the Scientific Committee for their input to the programme planning. Thanks are due to the Organizing Committee for their work during the preparatory phases and in the event itself. The Conference Secretariat, Jaana Lehtinen and Pihla Oksanen, took care of the website and the many practical and emergency arrangements. Maria Valaste edited the proceedings, just as one of her tasks. Jenna Wallenius handled the financial matters and Timo Harmo helped in IT issues. All these persons are of University of Helsinki. Alsuhail Faiz was responsible of arrangements in Statistics Finland. Several other people of the Department of Social Statistics and Statistics Finland assisted in practical arrangements during the conference. Thanks are due to all these persons. Last but not the least, the financial support kindly given by the International Association of Survey Statisticians, Nordic Council of Ministers, Federation of Finnish Learned Societies, SAS Institute, Statistics Finland and University of Helsinki, is greatly appreciated.

The year 2015 marks the 375th anniversary of the University of Helsinki and the 150th anniversary of Statistics Finland. Congratulations to both organizations!

I wish everybody an inspiring conference and enjoyable stay in Helsinki.

Risto Lehtonen

In Memoriam

Gunnar Kulldorff

06.12.1927 – 25.06.2015

The Baltic-Nordic-Ukrainian Network on Survey Statistics honours the memory of Gunnar Kulldorff, the founder and long-term chair of the network.

Gunnar was born in Malmö. He studied and also defended his PhD thesis "Estimation from Grouped and Partially Grouped Samples" in the University of Lund. He worked as a lecturer of statistics in the same University until 1965. After that he moved to Umeå where he lived the rest of his life. Gunnar became a professor in newly established Umeå University, first in statistics then in mathematical statistics. As first Dean of the Faculty of Philosophy and then of the Faculty of Mathematics and Natural Sciences he contributed a lot to building up Umeå University. Throughout his career he was much appreciated as teacher, researcher, leader and colleague.

Gunnar has always considered professional communication, exchange of scientific knowledge, cooperation and consolidation very important. International feature of the science was what he always stressed. As a member, often leader, of professional organizations, he has worked actively to strengthen these principles in the field of statistics. He was president of the International Statistical Institute (ISI) in 1989-1991; he has been chairman of Swedish Statistical Association, board member of the ISI, American Statistical Association and Bernoulli Society. He was elected as honorary member of the Finnish Statistical Society and the Estonian Statistical Society. In 2006 he has been awarded a degree of Doctor Honoris Causa by the University of Vilnius.

Gunnar was a great visionary. He has travelled to Asia, Africa and Mexico to initiate and support developments in statistics. However, most of his activity and energy has been directed to the countries not so far from his homeland. His great mission was to spread statistical culture, professional education and cooperation in newly re-independent Baltic Countries and later also in Ukraine and Belarus. His activity has created a vital network of survey statisticians that has now been functioning more than 20 years. In fact, Gunnar opened a door to the World for these countries. Before 1992, survey sampling was an unknown scientific field without any practical applications in these countries. Experienced countries of the Network, the Nordic countries, helped to overcome this shortage very soon. Survey Sampling courses were introduced to the curricula of the universities. More than 100 Bachelor and Master theses in this field were defended in the following years, as well as Doctoral theses. Survey practice started to develop in the National Statistical Agencies. Since 1997 annual workshops were organized in different Baltic countries, Ukraine and Belarus. These workshops formed a forum where students and teachers from the universities as well practitioners from the statistical agencies could present their results, put forward difficult problems and exchange experiences. Gunnar and Umeå University, hosting numerous visitors from the Network Countries, have played leading roles in educating new survey statisticians.

The Network will miss Gunnar's enthusiasm, personal warmth and sense of humour. Every single person in the Network has felt Gunnar's sincere interest and helpful attitude. Many people have enjoyed his concern

and help while visiting Umeå University. In the Network's conferences and annual workshops he was not merely present; he was the key figure. His organizational talent and personal appeal has created unforgettable unique atmosphere in these events. Memory of Gunnar remains in the hearts of people. The 4th Baltic-Nordic Conference on Survey Statistics (Helsinki 2015) confirms the vitality of the Network.

On behalf of the Baltic-Nordic-Ukrainian Network with gratitude and deep sadness,

Imbi Traat (Estonia)
Risto Lehtonen (Finland)

Contents

1 PROGRAMME 1

2 KEYNOTE PAPERS 9

Jelke Bethlehem: *Challanges of Web Surveys and Web Panels* 10

Jelke Bethlehem: *The Ever Changing Landscape of Statistical Data Collection* 11

Partha Lahiri: *Statistical Modeling and Estimation for Linked Data* 12

Partha Lahiri: *Can BIGDATA Help in the Production of Reliable Local Area Statistics?* 13

Eric Schulte Nordholt: *Combining Register and Survey Information in the Dutch Census 2011* 14

Eric Schulte Nordholt: *Statistical Disclosure Control Methods for Microdata in the Netherlands* 15

3 INVITED PAPERS 17

Per Gösta Andersson and Carl-Erik Särndal: *Reduced bias and increased variance: a possible trade-off in calibration for nonresponse treatment?* 18

Yves Berger: *Recent Advances in Empirical Likelihood Approaches under Complex Sampling* 19

Natallia Bokun: *Micro-entities survey design in Belarus* 20

Seppo Laaksonen: *Sampling design and weighting in the European Social Survey* 21

Yan Li: *Genetic Analyses Using Family-Based Survey Data* 23

Mārtiņš Liberts: *European health interview survey in Latvia — Challenges and Opportunities* 24

Inga Masiulaitytė-Šukevič: *Migration Statistics. Challenges for Statistics Lithuania* 25

Ulrich Rendtel, Marcus Gross, Timo Schmid, Sebastian Schmon and Nikos Tzavidis: *Bayesian kernel density estimation applied to sensitive geo-coded data of Berlin* 26

Bernardo Rota and Thomas Laitila: *Calibrating on Principal Components in the presence of Multiple Auxiliary Variables for Nonresponse Adjustment* 27

Kaija Ruotsalainen: *Register-based population census methodology in Finland* 28

Kaja Sõstra: *Use of administrative data for official statistics in Statistics Estonia* 29

Joonas Tuhkuri: *Big data: Google searches predict unemployment* 30

Ari Veijanen: *Effect of register errors on quality of survey estimates* 31

Ariane Würbach and Sabine Zinn: *Bayesian estimation of a general heaping model via different random-walk Metropolis specifications* 32

Li-Chun Zhang and John Dunne: *Census-like population size estimation based on administrative data* 33

4 CONTRIBUTED PAPERS 35

Manfred Antoni, Basha Vicari and Daniel Bela: *Interviewers' influence on bias in reported income* 36

Folasade Ariyibi and Salah Merad: *Improving contact rates in the field through analysis of linked Census survey data* 37

Maciej Beręsewicz: *A model-based approach to estimate bias in internet data sources* 38

Maciej Beręsewicz: *Assessing selectivity and representativeness of internet data sources for the real estate market in Poland* 39

Anastacia Bobrova, Iryna Andras and Andrei Piliutsik: *Sample survey of family to identify the intentions on having children* 40

Iana Bondarenko and Valery Turchyn: *Big data: one approach to processing ATM data* 42

Natalja Budkina and Mārtiņš Liberts: *On different points of view to the study of survey statistics* 43

Mariia Chebanova: *Aspects of Sampling Usage for Rare Populations for Labor Migration Measuring in Ukraine* 44

Ieva Dirdaitė: *Relationship between balanced sampling and calibrated estimator* 45

Kari Djerf, Atte Lintilä, Riku Salonen and Ari Veijanen: *Accuracy of imputation: a Case Study on the Finnish Labour Force Survey* 46

Kristina Galiautdinovaitė: *Income inequality measures for Baltic and Nordic countries* 47

Dan Hedlin: *Are we witnessing the end of random sampling in surveys?* 48

Tommi Härkänen, Juha Karvanen, Hanna Tolonen, Risto Lehtonen, Kari Djerf, Teppo Juntunen and Seppo Koskinen: *Systematic handling of missing data in complex study designs – experiences from the Health 2000 and 2011 Surveys* 49

Tetiana Ianevych: *Some approaches in analyzing the data with excess of zeros* 50

Anna-Kaisa Jaakkonen, Mika Kuoppa-aho and Johanna Laiho-Kauranne: *Improving efficiency of the sample design in the Finnish horticultural survey* 51

Aydin Karakoca and Alper Sinan: *Genetic modelling imputation approach for quantitative missing data problem* 52

Juha Karvanen, Ari Rantanen and Lasse Luoma: *Survey data and Bayesian analysis: a cost-efficient way to estimate customer equity* 53

Esa Katajamäki, Pasi Mattila and Johanna Laiho-Kauranne: *Reduction of response burden by replacing survey questions with register data: Cases of crop rotation, non-regular non-family labour, and number of animals* 54

Mauno Keto and Erkki Pahkinen: *Model-based optimal sample allocation for planned areas using EBLUP estimation* 55

Una Kojalo and Ģirts Briģis: *Trends in cervical cancer incidence in Latvija in 1983–2013* 56

Juho Kopra, Tommi Härkänen, Hanna Tolonen and Juha Karvanen: *Correcting for non-ignorable missingness in health indicator trends* 57

Danutė Krapavickaitė: *Small area estimation for a study variable having many zero values* 58

Anna Larchenko: *Statistical estimation and analysis of foreign trade in health services of the republic of Belarus* 59

Risto Lehtonen and Ari Veijanen: *Small area estimation by calibration methods* 60

Kaur Lumiste: *Auxiliary information in data collection and estimation stage* 61

Olha Lysa: *Combining HBS and LFS Data for Education Level Estimation* 62

Ethel Maasing: *First Results in Determining Permanent Residency Status in Register-Based Census* 63

Pentti Moilanen and Anssi Ahvonen: *The Fishing Management Fee Register and the Population Register as sampling frames in a Finnish recreational fishing survey* 64

Markku Mikael Nurminen: *Estimating the length of working careers from the Finnish labour force survey data* 65

Oona Pentala, Tommi Härkänen and Risto Kaikkonen: *Comparison of missing data methods using register-based auxiliary data for health-related survey data outcome* 66

Jaakko Reinikainen and Juha Karvanen: *Bayesian subcohort selection for longitudinal covariate measurements in follow-up studies* 67

Wojciech Roszka: *Synthetic data sources in the spatial analysis of poverty in Poland* 68

Iryna Rozora and Olga Lukovych: *Mean estimation with robust calibrated estimators* 69

Tomas Rudys: *An application of unit-level model for fractions of unemployed* 70

Volodymyr Sarioglo: *Improving of the Reliability of Ukrainian Poverty Indicators Estimates Using Auxiliary Information* 71

Rudi Seljak and Kaja Malešič: *Generalized solutions for data editing at SURS* 72

Alper Sinan and Aydın Karakoca: *A different imputation approach for the categorical survey studies* 73

Alina Sinisalo: *The development of production costs in dairy farms using panel data* 74

Alina Sinisalo, Arto Latukka and Anne-Mari Sepponen: *Testing differences of means* 75

Inga Skendere and Julija Stare: *Analysis of RSU International Brand — Views by International Students* 76

Markus Gintas Šova: *Challenges in Integrating Administrative VAT Data into UK Short-term Output Statistics* 77

Ene-Margit Tiit: *Residence testing using registers — conceptual and methodological problems* 78

Hanna Tolonen, Juha Karvanen, Päivikki Koponen, Erkki Vartiainen and Kari Kuulasmaa: *What can be learned about survey non-response through record linkage? Examples from health examination surveys* 79

Anton Tovchenko and Olexiy Tkachenko: *Outlier detection methods for buiseness surveys* 80

Maria Valaste: *Child care choices in Finland: coping with incomplete register-based data* 82

Olga Vasylyk: *A survey of student surveys* 83

Paavo Väisänen: *Metadata of the European Time Use Survey database* 84

Jacek Wesołowski: *An eigenproblem approach to optimal equal-precision sample allocation in subpopulations* 85

5 PARTICIPANTS 86

1 PROGRAMME

BaNoCoSS-2015
Fourth Baltic-Nordic Conference on Survey Statistics
24–28 August 2015 Helsinki

Programme

Updated 21 Aug. 2015

Monday 24 August Minerva Plaza

10:30–13:00	*Registration – Registration desk open at* Minerva Plaza
13:00–13:15	**Opening** Minerva Plaza Main Hall Risto Lehtonen
13:15–16:30	**PLENARY SESSIONS**
13:15–15:00	**Session 1** *Keynote talk* Minerva Plaza Main Hall *Chair* Risto Lehtonen **Partha Lahiri** (University of Maryland) Statistical modeling and estimation for linked data Discussion Can BIGDATA help in the production of reliable local area statistics? Discussion
15:00–15:30	*Refreshments – Registration desk open at* Minerva Plaza
15:30–16:30	**Session 2** *Invited papers* Minerva Plaza Main Hall *Chair* Olga Vasylyk
15:30–16:00	**Per Gösta Andersson** (Stockholm University) and **Carl-Erik Särndal (**Stockholm University**)**: Reduced bias and increased variance: a possible trade-off in calibration for nonresponse treatment? Discussion
16:00–16:30	**Ariane Würbach** (Leibniz Institute for Educational Trajectories & Otto-Friedrich-University Bamberg, Germany) and **Sabine Zinn** (Leibniz Institute for Educational Trajectories, Germany): Bayesian estimation of a general heaping model via different random-walk Metropolis specifications Discussion
16:40–17:40	**PARALLEL SESSIONS**
	Session 3 *Survey sampling & Estimation* Minerva Plaza Main Hall *Chair* Kaja Sõstra
16:40–17:00	**Ieva Dirdaitė** (Vilnius Gediminas Technical University): Relationship between balanced sampling and calibrated estimator
17:00–17:20	**Mariia Chebanova** (Taras Shevchenko National University of Kyiv): Aspects of sampling usage for rare populations for labor migration measuring in Ukraine
17:20–17:40	**Jaakko Reinikainen** (University of Jyväskylä) and **Juha Karvanen** (University of Jyväskylä): Bayesian subcohort selection for longitudinal covariate measurements in follow-up studies
	Session 4 *Survey Methodology & Analysis* Minerva Plaza Room K213 *Chair* Kaur Lumiste
16:40–17:00	**Iryna Andras** (Institute of Economics of NAS, Belarus), **Andrei Piliutsik** (Institute of Economics of NAS, Belarus) and **Anastacia Bobrova** (Institute of Economics of NAS, Belarus): Sample survey of family to identify the intentions on having children
17:00–17:20	**Kristina Galiautdinovaitė** (Vilnius Gediminas Technical University): Income inequality measures for Baltic and Nordic countries
17:20–17:40	**Aydin Karakoca** (Necmettin Erbakan University, Turkey) and **Alper Sinan** (Sinop University, Turkey): Genetic modelling imputation approach for quantitative missing data problem
17:50–18:30	**POSTER SESSION**
17:50–18:30	**Session 5** *Poster Session* Minerva Plaza Main Hall **Iana Bondarenko** (Oles Honchar Dnipropetrovsk National University, Ukraine) and **Valery Turchyn** (Oles Honchar Dnipropetrovsk National University, Ukraine): Big data: one approach to processing ATM data **Natalja Budkina** (Riga Technical University) and **Mārtiņš Liberts** (University of Latvia): On different points of view to the study of survey statistics **Markku Mikael Nurminen** (University of Helsinki): Estimating the length of working careers from the Finnish Labour Force Survey data **Alina Sinisalo** (Natural Resources Institute Finland, LUKE), **Arto Latukka** (LUKE) and **Anne-Mari Sepponen** (LUKE): Testing differences of means **Inga Skendere** (Riga Stradins University) and **Julija Stare** (Riga Stradins University): Analysis of RSU International Brand – Views by International Students
19:00–21:00	**University of Helsinki Reception** University Main Building (Fabianinkatu 33), Press Hall Foyer (2nd floor) Hosted by Dean Liisa Laakso

Tuesday 25 August Minerva Plaza	
9:00–13:50	**PLENARY SESSIONS**
9:00–10:30	**Session 6** *Invited papers* Minerva Plaza Main Hall *Chair* Imbi Traat
9:00–10:00	**Li-Chun Zhang** (University of Southampton and Statistics Norway) and **John Dunne** (Central Statistics Office, Ireland): Census-like population size estimation based on administrative data Discussion
10:00–10:30	**Ari Veijanen** (Statistics Finland): Effect of register errors on quality of survey estimates Discussion
10:30–11:00	*Refreshments*
11:00–12:00	**Session 7** *Survey Methodology* Minerva Plaza Main Hall *Chair* Kari Djerf
11:00–11:20	**Manfred Antoni** (Institute for Employment Research, Germany), **Basha Vicari** (Institute for Employment Research, Germany) and **Daniel Bela** (Leibniz Institute for Educational Trajectories): Interviewers' influence on bias in reported income
11:20–11:40	**Folaṣade Ariyibi** (Office for National Statistics, UK) and **Salah Merad** (Office for National Statistics, UK): Improving contact rates in the field through analysis of linked Census survey data
11:40–12:00	**Tommi Härkänen** (National Institute for Health and Welfare, THL, Finland), **Juha Karvanen** (University of Jyväskylä), **Hanna Tolonen** (THL), **Risto Lehtonen** (University of Helsinki), **Kari Djerf** (Statistics Finland), **Teppo Juntunen** (THL) and **Seppo Koskinen** (THL): Systematic handling of missing data in complex study designs - experiences from the Health 2000 and 2011 Surveys
12:00–13:00	Lunch Restaurant Olivia
13:00–13:50	**Session 8** *Invited paper* Minerva Plaza Main Hall *Chair* Dan Hedlin
	Yan Li (University of Maryland): Genetic analyses using family-based survey data Discussion
14:00–15:00	**PARALLEL SESSIONS**
	Session 9 *Health Surveys* Minerva Plaza Main Hall *Chair* Maria Valaste
14:00–14:20	**Una Kojalo** (Riga Stradins University) and **Ģirts Briģis** (Riga Stradins University): Trends in cervical cancer incidence in Latvija in 1983-2013
14:20–14:40	**Anna Larchenko** (Ministry of Foreign Affairs of the Republic of Belarus): Statistical estimation and analysis of foreign trade in health services of the Republic of Belarus
14:40–15:00	**Oona Pentala** (National Institute for Health and Welfare, THL, Finland), **Tommi Härkänen** (THL) and **Risto Kaikkonen** (THL): Comparison of missing data methods using register-based auxiliary data for health-related survey data outcome
	Session 10 *Small Area Estimation* Minerva Plaza Room K213 *Chair* Jacek Wesołowski
14:00–14:20	**Mauno Keto** (University of Jyväskylä) and **Erkki Pahkinen** (University of Jyväskylä): Model-based optimal sample allocation for planned areas using EBLUP estimation
14:20–14:40	**Danutė Krapavickaitė** (Vilnius Gediminas Technical University): Small area estimation for a study variable having many zero values
14:40–15:00	**Risto Lehtonen** (University of Helsinki) and **Ari Veijanen** (Statistics Finland): Small area estimation by calibration methods
15:00–15:30	*Refreshments*
15:30–16:30	**PLENARY SESSION**
	Session 11 *Invited papers* Minerva Plaza Main Hall *Chair* Tetiana Ianevich
15:30–16:00	**Mārtiņš Liberts** (University of Latvia): European Health Interview Survey in Latvia – challenges and opportunities Discussion
16:00–16:30	**Bernardo Rota** (Örebro University) and **Thomas Laitila** (Örebro University): Calibrating on principal components in the presence of multiple auxiliary variables for nonresponse adjustment Discussion
	Continued

16:40–18:10	PARALLEL SESSIONS
	Session 12 *Survey Statistics* Minerva Plaza Main Hall *Chair* Jyrki Möttönen
16:40–17:00	**Maciej Beręsewicz** (Poznan University of Economics): Assessing selectivity and representativeness of Internet data sources for the real estate market in Poland
17:00–17:20	**Kari Djerf** (Statistics Finland), **Atte Lintilä** (Statistics Finland), **Riku Salonen** (Statistics Finland) and **Ari Veijanen** (Statistics Finland): Accuracy of imputation: a Case Study on the Finnish Labour Force Survey
17:20–17:40	**Juho Kopra** (University of Jyvaskyla), **Tommi Härkänen** (National Institute for Health and Welfare, THL, Finland), **Hanna Tolonen** (THL) and **Juha Karvanen** (University of Jyvaskyla): Correcting for non-ignorable missingness in health indicator trends
17:40–18:00	**Alper Sinan** (Sinop University, Turkey) and **Aydın Karakoca** (Necmettin Erbakan University, Turkey): A different imputation approach for the categorical survey studies
	Session 13 *Business & Agricultural Surveys / Invited and contributed papers* Minerva Plaza Room K213 *Chair* Anna Larchenko
16:40–17:10	**Natallia Bokun** (Belarus State Economic University): Micro-entities sample survey design problems
17:10–17:30	**Esa Katajamäki** (Natural Resources Institute Finland, LUKE), **Pasi Mattila** (LUKE) and **Johanna Laiho-Kauranne** (LUKE): Reduction of response burden by replacing survey questions with register data: Cases of crop rotation, non-regular non-family labour, and number of animals
17:30–17:50	**Alina Sinisalo** (Natural Resources Institute Finland, LUKE): The development of production costs in dairy farms using panel data
17:50–18:10	**Markus Gintas Šova** (Office for National Statistics, UK): Challenges in Integrating Administrative VAT Data into UK Short-term Output Statistics

Wednesday 26 August	Nature Centre Haltia
9:00–10:00	Bus from Helsinki to Nature Centre Haltia Bus leaves from Kiasma bus stop, address: Mannerheiminaukio 2
10:00–11:30	**Session 14** *Keynote talk* Haltia Auditorium *Chair* Juha Alho **Jelke Bethlehem** (University of Leiden) Challenges of web surveys and web panels Discussion The ever changing landscape of statistical data collection Discussion
11:30–12:00	*Refreshments*
12:00–13:30 12:00–12:30 12:30–12:50 12:50–13:10 13:10–13:30	**Session 15** *Invited & contributed papers* Haltia Auditorium *Chair* Mārtiņš Liberts **Joonas Tuhkuri** (The Research Institute of the Finnish Economy ETLA): Big Data: Google searches predict unemployment Discussion **Maciej Beręsewicz** (Poznan University of Economics): A model-based approach to estimate bias in Internet data sources **Dan Hedlin** (Stockholm University): Are we witnessing the end of random sampling in surveys? **Kaur Lumiste** (University of Tartu): Auxiliary information in data collection and estimation stage
13:30–14:30	Lunch Restaurant Haltia
14:30–15:30	**Guided Haltia tour**
15:30–17:00	**Leisure time** **Lakeshore Sauna Session** Enjoy Finnish sauna and swimming in clear lake water! Sauna for females Sauna for males
17:00–18:00	Bus to Helsinki Bus leaves from Haltia parking place

Thursday 27 August	Statistics Finland
9:00–16:30	**PLENARY SESSIONS**
9:00–10:30	**Session 16** *Keynote talk* Auditorium 1 *Chair* Faiz Alsuhail **Eric Schulte Nordholt** (Statistics Netherlands): Combining register and survey information in the Dutch Census 2011 Discussion Statistical Disclosure Control methods for microdata in the Netherlands Discussion
10:30–11:00	*Refreshments*
11:00–12:00 11:00–11:30 11:30–12:00	**Session 17** *Invited papers* Auditorium 1 *Chair* Johanna Laiho-Kauranne **Kaja Sõstra** (Statistics Estonia): Use of administrative data for official statistics in Statistics Estonia Discussion **Ulrich Rendtel** (Freie Universität Berlin, FuB)**, Marcus Gross** (FuB), **Timo Schmid** (FuB)**, Sebastian Schmon** (FuB) and **Nikos Tzavidis** (University of Southampton): Bayesian kernel density estimation applied to sensitive geo-coded data of Berlin Discussion
12:00–13:00	Lunch Restaurant Stateria
13:00–13:50	**Session 18** *Invited paper* Auditorium 1 *Chair* Danutė Krapavickaitė **Kaija Ruotsalainen** *Invited paper* (Statistics Finland): Register-based population census methodology in Finland Discussion
14:00–15:00 14:00–14:20 14:20–14:40 14:40–15:00	**Session 19** *Administrative data & Census* Auditorium 1 *Chair* Ene-Margit Tiit **Ethel Maasing** (Statistics Estonia): First Results in Determining Permanent Residency Status in Register-Based Census **Wojciech Roszka** (Poznan University of Economics): Synthetic data sources in the spatial analysis of poverty in Poland **Hanna Tolonen** (National Institute for Health and Welfare, THL, Finland), **Juha Karvanen** (University of Jyväskylä), **Päivikki Koponen** (THL), **Erkki Vartiainen** (THL) and **Kari Kuulasmaa** (THL): What can be learned about survey non-response through record linkage? Examples from health examination surveys
15:00–15:30	*Refreshments*
15:30–16:30 15:30–15:50 15:50–16:10 16:10–16:30	**Session 20** *Sampling & Estimation* Auditorium 1 *Chair* Juha Karvanen **Tetiana Ianevych** (Taras Shevchenko National University of Kyiv): Some approaches in analyzing the data with excess of zeros **Tomas Rudys** (Vilnius University): An application of unit-level model for fractions of unemployed **Jacek Wesołowski** (Central Statistical Office, Poland): An eigenproblem approach to optimal equal-precision sample allocation in subpopulations
16:40–18:00	**PARALLEL SESSIONS**
 16:40–17:00 17:00–17:20 17:20–17:40 17:40–18:00	**Session 21** *Sampling & Estimation* Auditorium 1 *Chair* Markus Gintas Šova **Iryna Rozora** (Taras Shevchenko National University of Kyiv) and **Olga Lukovych** (Taras Shevchenko National University of Kyiv): Mean estimation with robust calibrated estimators **Pentti Moilanen** (Natural Resources Institute Finland, LUKE) and **Anssi Ahvonen** (LUKE): The Fishing Management Fee Register and the Population Register as sampling frames in a Finnish recreational fishing survey **Volodymyr Sarioglo** (National Academy of Science of Ukraine): Improving of the reliability of Ukrainian poverty indicators estimates using auxiliary information **Maria Valaste** (Social Insurance Institution, Finland):Child care choices in Finland: coping with incomplete register-based data
 16:40–17:00 17:00–17:20 17:20–17:40 17:40–18:00	**Session 22** *Survey Methodology & Analysis* Auditorium 2 *Chair* Kimmo Vehkalahti **Olha Lysa** (National Academy of Science of Ukraine): Combining HBS and LFS Data for Education Level Estimation **Rudi Seljak** (Statistical Office of the Republic of Slovenia) and **Kaja Malešič** (Statistical Office of the Republic of Slovenia): Generalized solutions for data editing at SURS **Olga Vasylyk** (Taras Shevchenko National University of Kyiv): A survey of student surveys **Paavo Väisänen** (Statistics Finland): Metadata of the European Time Use Survey database
18:00–19:00	
19:00–	**Conference Dinner** Restaurant Lasipalatsi (Mannerheimintie 22-24, Helsinki)

Friday 28 August	Statistics Finland
9:00–10:30	**Session 23** *Invited papers* Auditorium 1 *Chair* Tommi Härkänen
9:00–10.00	**Yves Berger** (University of Southampton): Recent advances in empirical likelihood approaches under complex sampling Discussion
10:00–10:30	**Seppo Laaksonen** (University of Helsinki): Sampling design and weighting in the European Social Survey Discussion
10:30–11:00	*Refreshments*
11:00–12:30	**Session 24** *Invited & contributed papers* Auditorium 1 *Chair* Ulrich Rendtel
11:00–11:30	**Inga Masiulaitytė-Šukevič** (Statistics Lithuania): Migration statistics. Challenges for Statistics Lithuania Discussion
11:30–11:50	**Juha Karvanen** (University of Jyväskylä), **Ari Rantanen** (Sanoma Media Finland) and **Lasse Luoma** (Tietoykkönen Oy, Finland): Survey data and Bayesian analysis: a cost-efficient way to estimate customer equity
11:50–12:10	**Anna-Kaisa Jaakkonen** (Natural Resources Institute Finland, LUKE), **Mika Kuoppa-aho** (LUKE) and **Johanna Laiho-Kauranne** (LUKE): Improving efficiency of the sample design in the Finnish horticultural survey
12:10–12:30	**Ene-Margit Tiit** (Statistics Estonia and University of Tartu): Residence testing using registers – conceptual and methodological problems
12:30–13:00	**Closing** Risto Lehtonen

2 KEYNOTE PAPERS

CHALLENGES OF WEB SURVEYS AND WEB PANELS

Jelke Bethlehem
Leiden University, The Netherlands, jelkeb@xs4all.nl

Traditionally, national statistical institutes collect data by means of face-to-face or telephone surveys. This is an expensive way of survey data collection, but experience has shown that it is necessary in order to produce high quality statistics. Nowadays, national statistical institutes in many countries are faced with budget constraints. This causes them to look for less expensive ways of data collection, while maintaining their quality. Web surveys seem a promising alternative. They have become increasingly popular, particularly in the world of market research. This is not surprising as a web survey is a simple, fast an inexpensive means to collect a lot of data.

At first sight, a web is just another mode of data collection. Questions are not asked face-to-face or by telephone, but over the internet. However, web surveys also suffer from methodological problems, such as under-coverage, self-selection, non-response and measurement errors. So, the question is if and how web surveys can be used for making official statistics.

One step further is a web panel. Once such a panel is in place and operational, a web survey can be conducted easy and fast. The sample can be selected from the list of panel members. Alternatively, all members can be approached. Only a questionnaire has to be designed and put on the internet. Then an e-mail is sent to all selected panel members. There are no costs involved. Response will be high, since all panel members agreed to participate in surveys regularly.

However, there is a caveat. Setting up a good, representative panel is no so easy. How to select a random sample (with equal probabilities) of people from the general population who want to become a member of the panel? And how to keep the web panel representative over time, as some panel members may lose interest in the course of time?

THE EVER CHANGING LANDSCAPE OF STATISTICAL DATA COLLECTION

Jelke Bethlehem
Leiden University, The Netherlands, jelkeb@xs4all.nl

The world needs statistics. All through history, there always have been statistics. And we will need statistics in the future too. But it will also become clear that the way in which we collect data for our statistics changes in the course of time.

The first part of the presentation is devoted to the past, and describes some important historic developments that led to the current situation. Among them are the rise of survey sampling and the increasing role of the computer for data collection.

The second part describes the current situation. The rapid rise of web surveys is discussed. Is this a good alternative for CAPI and CATI surveys? Other current challenges are reducing budgets for surveys, increasing nonresponse rates, and lack of proper sampling frames.

The third part of the presentation discusses possible future developments. Are there other means of data collection that avoid the current challenges? Several approaches are discussed:

- Non-probability sampling, including self-selection and sample matching.
- Model-based estimation
- Using big data
- Stick to survey sampling with improved correction techniques.

STATISTICAL MODELING AND ESTIMATION FOR LINKED DATA

Partha Lahiri
Joint Program in Survey Methodology
University of Maryland, College Park, U.S.A
plahiri@umd.edu

Computerized record linkage (CRL) methods are frequently used by government statistical agencies to quickly and accurately link two or more large files that contain information on the same individuals or entities using available information, which typically does not include unique, error-free identification codes. Because CRL utilizes already existing databases, it enables new statistical analysis without the substantial time and resources needed to collect new data. The possibility of errors in linkage causes problems for estimating the relationships between variables in the linked dataset. We will present a simple method to correct mismatch biases of standard estimators using an enhancement of the existing mixture models on measurements of the similarity among pairs of records to estimate probabilities used in calculating record linkage weights. We will report findings from a simulation study to compare the alternative estimators. This work is joint with Ms. Judith Law, PhD student, University of Maryland, College Park, USA.

CAN BIGDATA HELP IN THE PRODUCTION OF RELIABLE LOCAL AREA STATISTICS?

Partha Lahiri
Joint Program in Survey Methodology
University of Maryland, College Park, U.S.A
plahiri@umd.edu

The demand for various socio-economic and health statistics for small geographical areas is steadily increasing at a time when survey agencies are looking for ways to reduce costs to meet fixed budgetary requirements. In the current survey environment, the application of standard sample survey methods for small areas, which require large sample, is generally not feasible from the cost consideration. One of the key factors that lead to the success of small area methodology, which typically uses implicit or explicit models to combine survey and administrative data sources, is the availability of strong auxiliary variables. The accessibility of big data from different sources is now bringing new opportunities for statisticians to develop innovative small area methods. In this talk, I will discuss the hierarchical Bayesian methodology for exploiting BIGDATA in producing reliable local area statistics.

COMBINING REGISTER AND SURVEY INFORMATION IN THE DUTCH CENSUS 2011

Eric Schulte Nordholt
Statistics Netherlands, The Netherlands, e.schultenordholt@cbs.nl

Since the last census based on a complete enumeration was held in 1971, the willingness of the population to participate has fallen sharply. Statistics Netherlands no longer uses census questionnaires and has found an alternative to the traditional census in the register-based census, using only existing data. The Dutch 2011 Census tables were produced by combining available register and sample survey data. Additional to the technique of repeated weighting, for the 2011 Census the micro macro method (an IPF method) was used as a key method to obtain consistent tables based on survey data. The register-based census is cheaper and more socially acceptable than a traditional census. The table results of the Netherlands are not only comparable with earlier Dutch censuses, but also with those of the other countries in the 2011 European Census Round. More information can be found in the book Dutch Census 2011.

References

Dutch Census 2011, Analysis and Methodology, Statistics Netherlands, The Hague/Heerlen, November, 2014.
http://www.cbs.nl/en-GB/menu/themas/dossiers/historischereeksen/publicaties/publicaties/archief/2014/2014-dutch-census-2011-pub.htm?Languageswitch=on

STATISTICAL DISCLOSURE CONTROL METHODS FOR MICRODATA IN THE NETHERLANDS

Eric Schulte Nordholt
Statistics Netherlands, The Netherlands, e.schultenordholt@cbs.nl

At Statistics Netherlands different options exist to get access to microdata. In the eighties of last century hardly any access was given to researchers. Gradually the number of facilities has grown. Since the nineties of last century microdata are released as public use file (PUF) or microdata under contract (MUC). More detailed data can be analysed via the on-site and remote access facilities.
Special co-operation used to take place via remote execution but nowadays it has become more common to define so-called co-operation projects in which Statistics Netherlands works closely together with one or more external partners. Such projects only run if they are profitable for all partners involved. As different access facilities in the Netherlands have different confidentiality risks, different Statistical Disclosure Control (SDC) rules are applied for different facilities. The more detail in the data provided implies the stricter the access is organised.

References

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer and P.P. de Wolf, 2012. *Statistical Disclosure Control, Wiley Series in Survey Methodology*, Chichester, United Kingdom: Wiley.

3 INVITED PAPERS

REDUCED BIAS AND INCREASED VARIANCE: A POSSIBLE TRADE-OFF IN CALIBRATION FOR NONRESPONSE TREATMENT?

Per Gösta Andersson
Department of Statistics, Stockholm University, per.gosta.andersson@stat.su.se

Carl-Erik Särndal
Department of Statistics, Stockholm University

Several ways of using auxiliary information for calibrated weighting adjustment under survey nonresponse are presented. Information is often present at two levels, the population level and the sample level. The many options available in executing the calibration derive from several factors: One is the order in which the two sources of information enters into calibration, a choice of a bottom-up as opposed to a top-down approach. Another is whether the calibration should be carried out sequentially in two steps, or in one single step with the combined information. A third question is whether one can simplify the procedure, at no major loss of accuracy, by transcribing individual population auxiliary data from the register to the sample units only. We make a systematic list of the possibilities arising for calibration adjustment in this setting. An empirical study will illustrate the effects of different scenarios.

References

Brick, J.M. (2013). "Unit nonresponse and weighting adjustments: A critical review." *Journal of Official Statistics*, 29, 329-353.

Deville, J.C. (2000). "Generalized calibration and application to weighting for nonresponse." Compstat, *Proceedings in Computational Statistics, 14th symposium*, Utrecht; J.G. Bethlehem and P.G.M. van der Heijden (eds). Heidelberg; Physica Verlag, 65-76.

Kalton, G., and Kasprzyk, D. (1986). "The treatment of missing data." *Survey Methodology*, 12, 1-16.

Kott, P.S. (2006). "Using calibration weighting to adjust for nonresponse and coverage errors." *Survey Methodology*, 32, 133-142.

Kott, P.S., and Chang, T. (2010) "Calibration weighting to adjust for nonignorable unit nonresponse." *Journal if the American Statistical Association*, 105, 1265-1275.

Lundström, S., and Särndal, C.E. (1999). "Calibration as standard method for treatment of nonresponse." *Journal of Official Statistics*, 15, 305-327.

Särndal, C.E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester, UK: Wiley.

RECENT ADVANCES IN EMPIRICAL LIKELIHOOD APPROACHES UNDER COMPLEX SAMPLING

Yves G. Berger
University of Southampton, UK, Y.G.Berger@soton.ac.uk

The approach proposed by Berger and De La Riva Torres (2016) gives design-consistent estimators of parameters which are solutions of estimating equations (e.g. averages, totals, quantiles, correlation, (non)linear regression parameters). It can be used to construct confidence intervals without variance estimates. These confidence intervals are not based on the normality of the point estimator. Linearisation (e.g. Binder, 1983; Deville, 1999), re-sampling (jackknife or bootstrap) (e.g. Rao *et al*., 1992) or joint-inclusion probabilities are not necessary, even when the parameter of interest is not linear. Berger and De La Riva Torres's (2016) approach gives consistent confidence intervals even when the sampling distribution is skewed (e.g. with domains or with outlying values), or when linearisation gives biased variance estimates. The proposed approach can be used to estimate generalised regression parameters (e.g. logistic regression) and to test if they are significant, under a design-based approach (Oguz-Alper and Berger, 2014, 2015). The population level information is naturally taken into account, without the need of a calibration distance function (e.g. Deville and Särndal, 1992). The empirical likelihood approach is a design-based approach. A super-population model is not necessary. The empirical likelihood approach proposed by Berger and De La Riva Torres (2016) is different from the pseudoempirical likelihood approach (Chen and Sitter, 1999). The empirical likelihood approach of Berger and De La Riva Torres (2016) will be presented. Extensions to non-response (Berger, 2015) and nuisance parameters will be also covered (Oguz-Alper and Berger, 2014, 2015).

References

Berger, Y. G. (2015) Empirical likelihood confidence intervals in the presence of unit non-response. *60th session of International Statistical Institute, Rio de Janeiro, Brasil*, 3pp.

Berger, Y. G. and De La Riva Torres, O. (2016) An empirical likelihood approach for inference under complex sampling design. *To appear in the Journal of the Royal Statistical Society, Series B*, 22pp. URL http://dx.doi.org/10.1111/rssb.12115.

Binder, D. A. (1983) On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279–292.

Chen, J. and Sitter, R. R. (1999) A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, **9**, 385–406.

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

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

Oguz-Alper, M. and Berger, Y. G. (2014) Empirical likelihood confidence intervals and significance test for regression parameters under complex sampling designs. *Proceedings of the Survey Research Method Section of the American Statistical Association, Joint Statistical Meeting, Boston*, 10pp.

Oguz-Alper, M. and Berger, Y. G. (2015) *Profile empirical likelihood in the presence of nuisance parameters and population level information under unequal probability sampling*. Southampton: Southampton Statistical Sciences Research Institute http://eprints.soton.ac.uk/376699.

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992) Some recent work on resampling methods for complex surveys. *Survey Methodology*, **18**, 209–217.

MICRO-ENTITIES SAMPLE SURVEY DESIGN PROBLEMS

Natallia Bokun
Belarus State Economic University, e-mail: nataliabokun@rambler.ru

For the last years the growing number of small enterprises has motivated the development of specialized methodology and software for micro-entities sample survey. Survey object is the organization with number of employees less than 15 persons.

Nowadays, the National Statistical Committee of the Republic of Belarus together with Department of Statistics (BSEU) makes the preparatory work on the implementation of Micro-Entities Sample Survey (MS). In November 2014 a test sample survey was conducted; since 2015 MS is provided on a regular basis. The first results of MS indicated the appearance of significant organizational and methodological problems: non-responses, the need for localization of the sample, the presence of atypical units, need in a combination of statistical weighting methods, samples in small domains.

This talk on micro-entities sampling has fur parts:

1) Sampling Frames that incorporate two files of economic units: micro-entities and private farms.
2) Sample design; territorial stratified univariate and multivariate (multidimensional) samples are used.
3) Statistical weighting that includes three methods: traditional Horvitz-Thompson estimator and calibration (GREG- and SYN-estimators).

The results of trial calculations testing the first of methodological and software sampling were obtained in 2014 and 2015 years. The use of combination of univariate and multidimensional samples, different weighting methods will provide very reliable information over larger number of variables: employment, wages fund, revenues and others. However, standard errors, calculated by separate indicators in the context of different kind of activity at regional level are rather high. The improvement of representativeness in region weighting procedure can be complicated by use of auxiliary calibration estimators.

References

Bokun, N., Chernyshova, T (1997). Methods of sample surveys. Ministry of statistics and analysis of Belarus, Research Institute of Statistics.

Bokun, N., (2010). Problems of multidimensional samples in retail trade. Questions of statistics 3, 52-60.

Bokun, N. Micro-entities sample survey: problems of design, formation and usage / N. Bokun // Workshop of Baltic – Nordic – Ukrainian network on Survey Statistics, Tallinn, Estonia, August 25-28, 2014. – P. 25-31.

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

SAMPLING DESIGN AND WEIGHTING IN THE EUROPEAN SOCIAL SURVEY

Seppo Laaksonen
University of Helsinki, Seppo.Laaksonen@Helsinki.Fi

The European Social Survey (ESS) has aimed to control the sample designs used by specifying sampling guidelines that should have been followed in each participating country. The main requirements are the use of probability sampling and the achievement of a minimum effective sample size that is determined by ineligibility rate, nonresponse rate, inclusion probabilities and clustering effects. Several sampling strategies have been used over the first seven rounds, but in most countries the design has not changed substantially.

The sampling requirements are not always well satisfied. In this paper we present key problems observed until now, and their trends. The following questions especially are considered and statistics over six rounds presented: (i) ineligibility has been growing and becoming more complex, (ii) nonresponse is a worsening problem and varies inside any country as well, (iii) design effects due to clustering vary by cluster size and intra-class-correlation (ICC); (iv) design effects due to variation in selection probabilities depend mainly on the type of sampling frame used. Since round 2, the prediction of effective sample size has improved as it has been possible for most countries to use parameter estimates from the previous rounds (ICC, eligibility rate, response rate, coefficient of variation of selection probabilities), but apparent changes over time in these parameters have caused difficulties.

The second section of the paper is to describe what happens after the fieldwork relating to sampling. First, the sampling design data file is created following the specific template. This file consists of all the gross sample units and its variables include those that give opportunity to create sampling weights, to analyse the survey quality, and to estimate. Its most important characteristics, including sampling design variables and weights, will be finally merged together with the real survey variables at respondent level, and then the survey analysis is ready to begin.

The ESS public use file currently includes the two types of weights, the analysis weights whose sum is the number of the respondents of each country (representing 15+ years old residents) for each country, and the country size weight. Their product multiplied by 10000 gives an ordinary sampling weight. There exists the two types of analysis weights since 2013; (i) the weights based on the design assuming that nonresponse is ignorable, and (ii) the raking ratio weights that adjust for nonresponse and in-eligibility to some extent. We thus see that the weighting can be further improved but a drawback is that the current sampling design files are not reasonably good in any country although many countries could do it better.

Fortunately, the minimum level of the sampling design file is achieved in each country so that the design based weights can be created. But even the raking-ratio weights are calculated separately taking benchmarking margins from an external source. The best situation would be such a file that consists of these variables and in addition, of a number of micro and macro auxiliary variables that give opportunity for creating a more sophisticated weight such as it is an appropriate combination of both a response propensity weight and a calibration weight.

References

Deville, J-C. & Särndal, C-E. (1992) Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 376-382.

Laaksonen, S. (2015). Sampling design data file. *New Techniques and Technologies in Statistics (NTTS)*. European conference. Brussels 10-13 March. http://www.cros-portal.eu/sites/default/files//NTTS2015%20proceedings.pdf

Laaksonen, S. & Heiskanen, M. (2014). Comparison of Three Modes for a Crime Victimization Survey, Journal of Survey Statistics and Methodology 2 (4): 459-483 doi:10.1093/jssam/smu018

Le Guenne, J. & Sautory, O. (2005). CALMAR 2 : Une Nouvelle Version de la Macro Calmar de Redressment D'Échantillion Par Calage. http://vserver-insee.nexen.net/jms2005/site/files/documents/2005/327_1-JMS2002_SESSION1_LE-GUENNEC-SAUTORY_CALMAR-2_ACTES.PDF

Lynn, P. & Gabler, S. & Häder, S. & Laaksonen, S. (2007). Methods for Achieving Equivalence of Samples in Cross-National Surveys. *Journal of Official Statistics*, 27, 1, 107-124.

European Social Survey website (europeansocialsurvey.org) including the sampling guidelines for the round 6: http://www.europeansocialsurvey.org/docs/round6/methods/ESS6_sampling_guidelines.pdf

GENETIC ANALYSES USING FAMILY-BASED SURVEY DATA

Yan Li
University of Maryland at College Park, MD, USA, yli6@umd.edu

In population-based household health surveys and case-control studies that collect disease risk factors and DNA samples, for example, the National Health and Nutrition Examination Survey (NHANES) and The US Kidney Cancer Study (KCS), a case-control study, offer the opportunity to unbiasedly estimate genetic frequencies, test for Hardy-Weinberg equilibrium, or study genetic associations with prevalent health related conditions (phenotypes) in well-defined target populations. These types of genetic analyses are typically not populationbased in most genetic studies, as these studies consist of nonrepresentative samples of the target population. Household surveys, such as NHANES, however, collect population representative sample using multistage geographical cluster sampling, where at the last stage blood-related individuals are often selected from the same sampled households. These types of data presents challenges to conducting genetic analysis because observations are correlated due to two types of clustering with one induced by the geographical cluster sampling, and the other induced by biological inheritance among multiple participants within the same sampled household. Populationbased case-control studies, such as KCS, sample individuals at differential rates, which requires sample weights incorporated in analyses in order to obtain unbiased inferences. We have developed efficient statistical methods to address the cluster sampling and sample weighting effects on genetic inferences. The proposed methods are evaluated analytically and via Monte Carlo simulation studies, and illustrated using data from the Hispanic Health and Nutrition Survey, NHANES and KCS.

EUROPEAN HEALTH INTERVIEW SURVEY IN LATVIA – CHALLENGES AND OPPORTUNITIES

Mārtiņš Liberts
Central Statistical Bureau of Latvia, martins.liberts@csb.gov.lv

The last European Health Interview Survey was done in 2014. The survey in Latvia was organised by the Central Statistical Bureau of Latvia. Several innovations were implemented for the sample design with an aim to improve the precision of the survey results. The population frame was constructed as a list of residents living in private households in age group 15 years and more. The frame was split into two primary strata – residents with phone number available (57%) and residents with phone number not available (43%). Different sampling designs and data collection modes were used for each stratum. Stratified systematic one-stage random sampling and CATI data collection was used for the first primary stratum. Stratified systematic two-stage random sampling and CAPI data collection was used for the second primary stratum. Administrative data from National Health Service were used to create an auxiliary variable for the population frame. The auxiliary variable is a dummy variable indicating residents who have received state paid health services during 2013. The auxiliary variable was used for stratification and it will be used also for the weighting of the survey data.

References

European Commission. (2013). European Health Interview Survey (EHIS wave 2). Luxembourg. Retrieved from http://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-RA-13-018

MIGRATION STATISTICS. CHALLENGES FOR STATISTICS LITHUANIA

Inga Masiulaitytė-Šukevič
Statistics Lithuania, Lithuania, inga.masiulaityte@stat.gov.lt

Many countries are facing demographic challenges manifested not only by declining fertility rate and rapidly ageing population, but also by migration processes. Dynamics and complexity of migration processes influence the formation of migration policy; therefore, over the recent years statisticians have drawn considerable attention to estimate emigration process (or its volume). In recent years, the international statistical community has intensified its efforts to improve availability, quality and comparability of available data on international migration. These efforts often include collection, estimation as well as dissemination of international migration statistics by relevant demographic and socio-economic characteristics.

Particular attention is given to the exhaustive analysis of administrative data sources with the aim to use those data for producing comprehensive return migration statistics. Population registers data and 2011 Population and Housing Census results are also used as a data source for the producing and development of return migration statistics.

The efforts and work undertaken by Statistics Lithuania on improving and developing international migration statistics in current globalization world will be presented.

References

Dumont J.-C., Spielvogel G. (2008), Return Migration: A New Perspective. OECD.

Jolivet M., Xenogiani T., Dumont J.-C. (2012), Measuring return migration: some preliminary findings in time of crisis. OECD. Economic Commission for Europe. Conference of European Statisticians.

Lapėnienė V. (2009), New approach to international migration statistics in Lithuania. Combination of data from labour force survey and population registers. Statistics Lithuania.

Lapėnienė V., Masiulaitytė-Šukevič I. (2014), Migration Statistics. Challenges for Statistics Lithuania. Statistics Lithuania. Economic Commission for Europe. Conference of European Statisticians.

National Statistics Institute of Spain (2012), Statistical approach to the migratory phenomenon in Spain during the period 2001-2011 from the recorded information in the Municipal Population Registers. Economic Commission for Europe. Conference of European Statisticians.

Statistics Netherlands (2012), Return migration rates of recent immigrants compared to flows in the previous century. Economic Commission for Europe. Conference of European Statisticians.

Sipavičienė A., Gaidys V., Dobrynina M. (2009), Return migration: theoretical insights and the situation in Lithuania. Vilnius, Lithuania.

Tegsjö B. (2005), Experience and proposals from a migration statistics project at Statistics Sweden. 28th CEIES seminar. Migration statistics. p. 123-131.

United Nations (1998), Recommendations on Statistics of International Migration. Revision 1, New York.

United Nations (2011), Statistics on international migration. A practical guide for countries of Eastern Europe and Central Asia.

United Nations (2014), Migration statistics. Report of the Secretary-General. Economic and Social Council.

BAYESIAN KERNEL DENSITY ESTIMATION APPLIED TO SENSITIVE GEO-CODED DATA OF BERLIN

Ulrich Rendtel
Freie Universität at Berlin, Germany, ulrich.Rendtel@fu-berlin.de

Marcus Gross
Freie Universität at Berlin, Germany, Marcus.Gross@fu-berlin.de

Timo Schmid
Freie Universität at Berlin, Germany, Timo.Schmid@fu-berlin.de

Sebastian Schmon
Freie Universität at Berlin, Germany, Sebastian.Schmon@fu-berlin.de

Nikos Tzavidis
University Southampton, UK, N.Tzavidis@soton.ac.uk

Modern systems of social statistics require the timely estimation of area-specific densities of sub-populations. Ideally estimates should be based on precise geo-coded information, which is not available due to confidentiality constraints. One approach for ensuring confidentiality is by rounding the geo-coordinates. We propose multivariate non-parametric kernel density estimation that reverses the rounding process by using a Bayesian measurement error model. The methodology is applied to the Berlin register of residents for deriving density estimates of ethnic minorities and aged people. Estimates are used for identifying areas with a need for new advisory centres for migrants and infrastructure for older people.

References

Carroll, R., D. Ruppert, L. Stefanski, and C. Crainiceanu (2006). Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Scott, D. W. and S. J. Sheather (1985). Kernel density estimation with binned data. Communications in Statistics - Theory and Methods 14 (6), 1353-1359.

Wang, B. and W. Wertelecki (2013). Density estimation for data with rounding errors. Computational Statistics & Data Analysis 65, 4-12.

CALIBRATING ON PRINCIPAL COMPONENTS IN THE PRESENCE OF MULTIPLE AUXILIARY VARIABLES FOR NONRESPONSE ADJUSTMENT

Bernardo Rota
Örebro University, Sweden, bernardo.rota@oru.se

Thomas Laitila
Örebro University, Sweden, thomas.laitila@oru.se

A prerequisite for valid estimation in surveys with nonresponse is access to appropriate auxiliary information. When a large set of auxiliary variables is available, estimating on all of these may result in inefficient estimators, especially if the set contains duplications or variables highly correlated. Thus, a subset of available auxiliary variables has to be selected. This selection has to be made with care avoiding exclusion of auxiliary variables bringing information on the estimation problem at hand. In this paper the principal components method is suggested for dimension reduction. The effectiveness of using principal components in two different calibration schemes is studied: the linear calibration that uses no explicit response function and the propensity calibration which is based on an explicit functional form. Furthermore, a principal component retention criteria based on the canonical correlation between the principal components and the model variables is suggested. Simulation results illustrate that the properties of the estimators are improved by using principal components of the auxiliary variables.

References

Bardsley, P. and Chambers, R. L. (1984). Multipurpose estimation from unbalanced samples.*Journal of the Royal Statistical Society.Series C(Applied Statistics)*,3:33,290-299.
Beaumont, J. F. (2006). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, **31**, 227-231.
Bethlehem, J. and B. Schouten (2004). Nonresponse adjustment in household surveys, *Discussion paper 04007. Statistics Netherlands, Voorburg/Heerlen, The Netherlands*.
Bilen, C., Khan, A. and Yadav, O. P. (2010). Principal components regression control for multivariate autocorrelated cascade process. *Int.J.Quality Engeneering and Technology*, **1:3**
Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279-292.
Brick, J. M. (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of Official Statistics*, **29:3**, 329-353
Borga, M. (2001). *Canonical Correlation a Tutorial.*
Retrieved from https://www.cs.cmu.edu/~+tom/10701 sp11/slides/CCA tutorial.pdf
Cadima, J. and Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principle components. *Journal of Applied Statistics*, **22:2**, 203-214, DOI: 10.1080/757584614
Cardot, H., Goga, C. and Shehzad, M.-A. (2014). Calibration and Partial Calibration on Principal Components when the Number of Auxiliary Variables is Largear. *Xiv:1406.7686v2 [stat.ME]*
Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model, *Biometrika*, **95:3**, 555-571.
Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376-382.
Deville, J.C. and Särndal, C. E. and Sautory, O. (1993). Generalized raking procedures in survey sampling, *Journal of the American Statistical Association*, **88:423**, 1013-1020.

REGISTER-BASED POPULATION CENSUS METHODOLOGY IN FINLAND

Kaija Ruotsalainen
Statistics Finland, Finland, Kaija.Ruotsalainen@stat.fi

In Finland the use of the administrative data and registers already has a history of over 40 years. The decisive step towards a register-based population system was taken at the end of the 1960s when the Central Population Register was established. By means of this system an identifying personal code was issued to each resident in Finland. The same personal identity number (PIN) was taken into use in other administrative registers, such as in taxation and in the employment pension insurance system.
The primary reasons for starting to exploit administrative and other register data are the lowering the costs, advancement in data systems and processing methods, reduction of response burdens and increased demand for information.
One of the main advantages of register-based statistical systems is that total data can be produced annually. Then the data collecting and processing systems are continuously maintained and updated, and the statistical data is quite timeliness compared e.g. census data, which is usually available every 10th year.
The use of registers for statistical purposes is by no means free from problems. One big challenge is dependence on data suppliers. An amendment in legislation or some other administrative change can cause changes to the con-tent of the registers as well. Other problems to be mentioned are e.g. that the coverage of registers may be defective for some data, although the register itself would contain all the units to be described. Also, there may arise consistency problems when linking information from different registers.
In Finland, utilisation of administrative data as a source for statistics in order to rationalise statistics production is accepted. This becomes particularly clear from the Finnish Statistics Act, according to which statistical authorities must primarily exploit existing data sources and the authorities controlling these registers are obliged to supply this data to statistical authorities.

References

Myrskylä, Pekka and Kaija Ruotsalainen (1997). Annual System of Small Area Statistics Based on Administrative Records and Registers. Contributed Papers for the 51st Session of the ISI, Istanbul. Book 1, 511-512.

Ruotsalainen, Kaija (2001). Annual System of Small Area Statistics Based on Administrative Records and Registers - the Possibilities and the Problems. A paper prepared for the forty-ninth plenary session of the Conference of European Statisticians, Geneva, 11-13 June 2001

Ruotsalainen, Kaija (2004). Use of Administrative Data in Population Censuses in Finland. A paper prepared for TACIS Seminar, Paris, 4-6 October 2004.

Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland, Statistics Finland, 2004

USE OF ADMINISTRATIVE DATA FOR OFFICIAL STATISTICS IN STATISTICS ESTONIA

Kaja Sõstra
Statistics Estonia, Estonia, kaja.sostra@stat.ee

The use of administrative data for producing statistics is increased over the last 5-8 years in Statistics Estonia. The Official Statistics Act of Estonia (2010) foresees that Statistics Estonia shall primary use data collected in administrative databases, if such data allow the production of official statistics complying with the quality criteria of official statistics. Several projects has started for developing methodology for more effective use of administrative data.

Estonia plans to conduct the next population and housing census (PHC) completely register-based. The preparations for the register-based PHC began in 2010 with the methodology project to check the quality and interoperability of state registers. The small-scale pilot census was conducted in 2014 using three registers and PHC 2011 data, two pilot censuses are planned in 2016 and 2018. Seventeen administrative data registers will be directly used for compiling PHC variables and some additional registers will be used for quality checks. One key problem of register-based PHC is the quality of place of residence in administrative sources. Special project for determining permanent residency status based on representation of persons in ten administrative registers has carried out.

The project for using electronic annual report (AR) data started in 2009 including technical and methodological preparation for receiving and using AR data. Statistics Estonia collects data from economic units (enterprises, institutions, organisations etc.) with monthly, quarterly and annual questionnaires. Some of these questionnaires include indicators that economic units can also submit in their annual report which data are submitted electronically to the Commercial Register. Since the AR does not include all necessary variables for compiling annual statistics the AR data is used in two ways: pre-filling of statistical questionnaires (starting from 2012) and cut-off sampling of small enterprises together with modelling missing variables (2014). As a result, the response burden of respondents has decreased.

The project for development of a methodological and statistical basis for a new indicator of services production was started in 2014. The main aim is to develop of collection and estimation methods for the production of monthly service turnover and volume indicator with a minimum increase in the reporting burden on enterprises. The idea is to use monthly VAT data and quarterly statistical survey data to predict monthly turnover and produce monthly indicators.

The use of administrative data for producing official statistics has advantages and problems compared to traditional surveys. The comprehensive methodological work is needed for solving problems: missing variables and values, differences in definitions, coverage problems etc. Nevertheless Statistics Estonia continues the work with better use of administrative data.

References

1. Matteus, D. (2013) Roadmap to register-based census. Quarterly Bulletin of Statistics Estonia 4/2013. ISSN-L 1736-7921; ISSN 1736-7921 (hard copy); ISSN 2346-6049 (PDF) http://www.stat.ee/dokumendid/75154
2. Official Statistics Act of Estonia (2010) https://www.riigiteataja.ee/en/eli/ee/506012015002/consolide/current
3. Tamm, E., Põldsaar, M., Nestor, R. (2015) Use of annual reports and statistical models in the production of statistics. Quarterly Bulletin of Statistics Estonia 2/2015. ISSN-L 1736-7921; ISSN 1736-7921 (hard copy); ISSN 2346-6049 (PDF) http://www.stat.ee/publicationdownload-pdf?publication_id=39431

BIG DATA: GOOGLE SEARCHES PREDICT UNEMPLOYMENT

Joonas Tuhkuri
ETLA, The Research Institute of the Finnish Economy and University of Helsinki, Finland,
joonas.tuhkuri@etla.fi

There are over 100 billion searches on Google every month. This paper examines whether Google search queries can be used to predict the present and the near future unemployment rate in the US. Predicting the present and near future is of interest, as the official records of the state of the economy are published with a delay. To assess the information contained in Google search queries, the paper compares a simple predictive model of unemployment to a model that contains a variable, Google Index, constructed from Google data. In addition, descriptive cross-correlation analysis and Granger non-causality tests are performed. To study the robustness of the results, the paper considers state-level variation in the unemployment rate and Google Index using a fixed effects model. Furthermore, the sensitivity of the results is studied with regard to different search terms. The results suggest that Google searches contain useful information on the present and the near future unemployment rate. The value of Google data for forecasting purposes, however, tends to be time specific, and the predictive power of Google searches appear to be limited to short-term predictions. The results demonstrate that big data can be utilized to forecast economic indicators.

EFFECT OF REGISTER ERRORS ON QUALITY OF SURVEY ESTIMATES

Ari Veijanen
Statistics Finland, ari.veijanen@stat.fi

Consider the estimation of the means of a survey variable in categories of a register variable. In Finland, unique personal identification codes can be used to obtain the values of register variables for a person in the sample. Registers are usually perceived as reliable but error rates of ten percent or more have been observed (Wallgren and Wallgren, 2014). Effects of misclassification have been studied by Zhang and Fosen (2012), for example. This paper describes how errors in register variables affect statistics as compared with results obtained using true values, which are unknown in practice.

The regional means of a survey variable are often estimated with the help of a model fitted to the sample. What happens when some of the explanatory register variables contain errors? Theory on errors-in-variables and measurement errors shows that the estimate of a slope parameter associated with an explanatory variable containing errors tends asymptotically to zero, as the variance of the error increases. It is not known how errors in auxiliary variables affect calibration, generalized regression estimator (GREG), or empirical best linear unbiased predictor (EBLUP). In a simulation experiment with a synthetic population, a large value was added to an auxiliary variable in the population with probability 0.01. Model-free domain-level calibration was sensitive to this contamination. Model calibration (Wu and Sitter 2001; Lehtonen and Veijanen, 2012), which involves predictions from a model instead of auxiliary variables, was much less sensitive. The mean squared error of GREG and EBLUP increased only slightly due to contamination. GREG was still design unbiased, whereas the design bias of EBLUP was affected by contamination in small domains. All these methods incorporated a mixed model.

Consider the class means of a survey variable Y, when the classification of units is obtained from a register. Because of misclassification, the available classification C' sometimes differs from the unknown true class C. When the sample size increases, even a design unbiased class mean estimator for class c tends to the expectation $E\left(Y \mid C' = c\right)$ given the classification, not to the true expectation $E\left(Y \mid C = c\right)$. The bias does not vanish. Under certain conditions, the bias can be approximated using

$$\left|E\left(Y \mid C' = c\right) - E\left(Y \mid C = c\right)\right| \leq \left(1 - P\{C = c \mid C' = c\}\right) \max_i \left|\mu_i - \mu_c\right|,$$

where $\mu_i = E\left(Y \mid C = i\right)$.

Typical sources of errors in a register are coding errors in auxiliary variables and misclassification due to a delayed update.

References

Lehtonen, R. & Veijanen, A. (2012). Small area poverty estimation by model calibration. Journal of the Indian Society of Agricultural Statistics, 66, 125-133.

Wallgren, A. & Wallgren, B. (2014). Register-based statistics: statistical methods for administrative data. 2nd edition.

Wu, C. & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association, 96, 185–193.

Zhang, L.-C. & Fosen, J. (2012). A modeling approach for uncertainty assessment of register-based small area statistics. Journal of the Indian Society of Agricultural Statistics, 66, 91-104.

BAYESIAN ESTIMATION OF A GENERAL HEAPING MODEL VIA DIFFERENT RANDOM-WALK METROPOLIS SPECIFICATIONS

Ariane Würbach
Leibniz Institute for Educational Trajectories (LIfBi), Germany, ariane.wuerbach@lifbi.de
Chair of Statistics and Econometrics, Otto-Friedrich-University Bamberg, Germany, ariane.wuerbach@uni-bamberg.de

Sabine Zinn
Leibniz Institute for Educational Trajectories (LIfBi), Germany, sabine.zinn@lifbi.de

Heaping is a general tendency of survey respondents to report income data rounded off to the nearby modulos. There can be differences to which degree heaping occurs, i.e. possible modulos can be 100, 500, or 1000. We developed a general method to account for different heaping patterns [6,5]. A mixture model describes the underlying zero-inflated log-normal distribution and the supposed heaping mechanism. For the sake of simplicity, we assume that the probabilities of heaping to specific heaping points (modulos) are constant within predefined intervals. The parameters of the mixture model are estimated simultaneously using five different random-walk Metropolis schemes for computation: a single-block scheme, a multiple-block scheme (MB), a randomized multiple-block scheme (RMB) [1], and two adaptive schemes (AP, AM) [3,4]. The results are compared by their inefficiency factors and marginal likelihoods [1,2]. Results from a simulation study show that estimates are better approximated by either MB or RMB schemes. The proposed method is applied to income data of the National Educational Panel Study (NEPS). The performance of application is explored by posterior predictive checks and demonstrates a good fit of the general heaping model.

Keywords: heaping; random-walk Metropolis algorithm; multiple-block scheme; adaptive MCMC

References

[1] Chib, S. and Ramamurthy, S. (2010): Tailored randomized block MCMC methods with application to DSGE models. *Journal of Econometrics*, 155(1), pp. 19–38.
[2] Frühwirth-Schnatter, S. (2006): Finite mixture and Markov switching models, Chapter 5.3. New York: Springer.
[3] Haario, H., Saksman, E. and Tamminen, J. (1999): Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14(3), pp. 375–395.
[4] Haario, H., Saksman, E. and Tamminen, J. (2001): An adaptive Metropolis algorithm. *Bernoulli*, 7(2), pp. 223–242.
[5] van der Laan, J. and Kuijvenhoven, L. (2011): Imputation of rounded data. *Statistics Netherlands*. The Hague/Heerlen (Discussion Paper, 201108).
[6] Zinn, S. and Würbach, A. (2014): A Statistical Approach to Account for Heaping Patterns: An Application to Self-Reported Income Data. *Leibniz Institute for Educational Trajectories (LIfBi). German National Educational Panel Study (NEPS)*. Bamberg (NEPS Working Papers, 40).

CENSUS-LIKE POPULATION SIZE ESTIMATION BASED ON ADMINISTRATIVE DATA

Li-Chun Zhang
University of Southampton, UK, L.Zhang@soton.ac.uk
Statistics Norway, Norway, lcz@ssb.no

John Dunne
Central Statistics Office, Ireland, John.Dunne@cso.ie

Register data that originate from administrative sources are increasingly being explored to generate statistical outputs directly. Using register data to produce census-like population size estimates is an on-going development at the Central Statistics Office, Ireland. It aims to reduce the cost associated with the traditional approach based on census and coverage surveys, making it feasible to update the population estimates on a yearly basis. The methodology can potentially provide a viable approach and a pioneering example for provision of population statistics, in a setting where one neither has a central population register nor deploys any additional coverage surveys.

Traditionally capture-recapture models have been used to deal with multiple list enumerations subjected to under-enumeration errors. A number of assumptions need to be checked when applying the method to the available register data, including independence between the lists, homogeneous capture probability, etc. In particular, compared to census and coverage surveys, the registers may contain non-negligible erroneous enumerations, or over-coverage error, which is a challenge to the existing capture-recapture estimation methods.

In this talk we discuss two recent developments in handling capture-recapture data with additional over-coverage errors. First, several alternative modelling assumptions regarding the erroneous enumerations are considered in details for the setting with two (or more) list enumerations that may suffer from erroneous enumerations and an additional list with only under-coverage error. This helps to determine, in a given data setting, which assumption is most useful. Next, we consider a trimmed dual-system estimation (TDSE) approach, which consists of scoring the potential erroneous records in the two lists on which the standard DSE is based, and apply the DSE after removing the flagged records. We discuss how stopping rules of the trimming can be constructed theoretically, and how the TDSE can be combined with the modelling approach, in order to verify the extent of the remaining erroneous enumerations and, if necessary, to accommodate them in the estimation.

4 CONTRIBUTED PAPERS

INTERVIEWERS' INFLUENCE ON BIAS IN REPORTED INCOME

Manfred Antoni
Institute for Employment Research (IAB), Germany, manfred.antoni@iab.de

Basha Vicari
Institute for Employment Research (IAB), Germany, basha.vicari@iab.de

Daniel Bela
Leibniz Institute for Educational Trajectories (LIfBi), Germany, daniel.bela@lifbi.de

We investigate characteristics of respondents and interviewers influencing the accurateness of reported income by comparing survey data with administrative data. Questions on sensitive topics like respondents' income often produce relatively high rates of item nonresponse or measurement error. In this context several analyses have been done on item nonresponse (e.g. Essig/Winter 2009), but little is known about misreporting. Existing evidence shows that it is unpleasant for respondents to report very low or very high income. The generally observed high rates of item nonresponse at both tails of the income distribution support this hypothesis (Riphahn/Serfling 2005).
One possible explanation of such misreporting is social desirability bias, which may lead to overreporting of desirable attributes or underreporting of undesirable ones in order to present oneself in a positive light (Stocké/Hunkler 2007). Because an experienced and competent interviewer may be able to inhibit such behavior, interviewer characteristics as well as their interaction with respondent characteristics should be of particular importance. Moreover, the bias should decrease with perceived social closeness between respondent and interviewer (Diekmann 2008).
Using linked survey and administrative data we are able to measure the extent of deviation between reported and recorded incomes and explore the influence of respondent and interviewer characteristics on it. The starting point for the linkage is data from the German National Educational Panel Study (NEPS), Starting Cohort 6 (see Allmendinger et al. 2011), which surveys adults from birth cohorts 1944 to 1986. More than 90% of the respondents consented to a linkage of their survey information with administrative data from the German Federal Employment Agency. These longitudinal earnings data are highly reliable as they are based on mandatory notifications of employers to the social security system.
We include interviewer and respondent characteristics as well as their interactions into our model to estimate their respective impact on the incidence and size of any bias in reported incomes. This allows us to control for latent interviewer traits that might have influenced the respondent's answering behavior during each interview of a given interviewer.

References

Allmendinger, J.,C. Kleinert, M. Antoni, B. Christoph, K. Drasch, F. Janik, K. Leuze, B. Matthes, R. Pollak, Ruland, M. (2011). “Adult education and lifelong learning”. In: Zeitschrift für Erziehungswissenschaft 14.Supplement 2, pp. 283–299.
Diekmann, A. (2008). Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen. 19. Reinbek bei Hamburg: Rowohlt.
Essig, L., Winter, J. (2009). “Item Non-Response to Financial Questions in Household Surveys: An Experimental Study of Interviewer and Mode Effects”. In: Fiscal Studies 30.4, pp. 367–390.
Riphahn, R., Serfling, O. (2005). “Item Non-Response on Income and Wealth Questions”. In: Empirical Economics 30.2, pp. 521–538.
Stocké, V., Hunkler, C. (2007). “Measures of Desirability Beliefs and Their Validity as Indicators for Socially Desirable Responding”. In: Field Methods 19.3, pp. 313–336.

IMPROVING CONTACT RATES IN THE FIELD THROUGH ANALYSIS OF LINKED CENSUS SURVEY DATA

Folaṣade Ariyibi
Office for National Statistics, UK, fola.ariyibi@ons.gsi.gov.uk

Salah Merad
Office for National Statistics, UK, salah.merad@ons.gsi.gov.uk

The UK Census is carried out every 10 years, with the last having taken place in 2011. As it is compulsory, the Census achieves very high responses rates, with the last Census obtaining a 94% response rate. This would allow analysis of non-response in survey data when linked with Census records, which we refer to as the Census Non-response Link Study (CNRLS). This has been carried out at every Census since 1991 by the Office for National Statistics.

This paper presents some of the analyses performed using data from the 2011 CNRLS, the aim of which is to improve data collection in the field. The analyses make use of Census data as well as paradata collected by field interviewers (i.e. time of day, number of times interviewer called). The results are then used to identify characteristics that can be observed in the field to inform how best to contact households.

A MODEL-BASED APPROACH TO ESTIMATE BIAS IN INTERNET DATA SOURCES

Maciej Beręsewicz
Poznań University of Economics, Poland, maciej.beresewicz@ue.poznan.pl

New data sources, such as the Internet data sources (IDS) or big data, are gaining recognition not only in the non-statistical but also statistical literature. Several papers and presentations devoted to these new data sources address the problem of representativeness, bias, data quality or, generally, the usefulness of these data for official statistics. Information from those sources is not only used for producing statistics, but is also increasingly being used in the model-based approach as a source of auxiliary variables. This is also true of small area estimation, particularly in the field of area-level models, for example supplied with data from Google Trends (Porter et al., 2014; Rao, 2014). Therefore, it is crucial to evaluate and quantify the bias that can be observed in these data sources.

In view of the above, the main aim of the paper is to present an attempt to indirectly estimate bias of selected characteristics (the price of m^2, number of rooms and floor area) of flats offered to sale on the secondary real estate market in Poland using information from IDS. For this purpose, the author will apply the model-based approach to estimate the bias of statistics based on administrative registers proposed by Zhang (2012). In addition, the paper will present an organization of statistical concepts related to Internet data sources and demonstrate the importance of assessing bias, which tends to be neglected in the non-statistical literature.

Furthermore, an extension of the concept of modelling bias described by Zhang (2012) will be discussed. The proposed approach will adapt a more general linear mixed model that takes into account (i) several data sources, (ii) domain and time series data and (iii) the context of new data sources, in particular IDS. The methods presented in the paper will be exemplified using actual data from real estate portals in Poland. All calculations will be conducted in the **R** statistical package (R Core Team, 2014).

The paper is the part of the research is supported by the National Science Centre, Poland, Preludium 7 grant no. 2014/13/N/HS4/02999.

References

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Porter, A. T., Holan, S. H., Wikle, C. K., & Cressie, N. (2014). Spatial Fay–Herriot models for small area estimation with functional covariates. Spatial Statistics, 10, 27-42.

Rao J.N.K. (2014), Inferential Issues in Model-based SAE: Some New Developments, International conference on SAE 2014, Poznań.

Zhang L.-C. (2012), On the accuracy of register-based census employment statistics, Presentation on European Conference on Quality in Official Statistics in Greece, 2012.

ASSESSING SELECTIVITY AND REPRESENTATIVENESS OF INTERNET DATA SOURCES FOR THE REAL ESTATE MARKET IN POLAND

Maciej Beręsewicz
Poznań University of Economics, Poland, maciej.beresewicz@ue.poznan.pl

Estimation conducted by National Statistical Institutes heavily relies on statistical data, such as surveys, census or establishments reporting. Recently, in an effort to meet information needs, official statisticians have been searching for new (non-statistical) data sources, including both government (administrative records) and non-government (private) data. The main examples of the latter group are Internet data sources (IDS) or big data, which have recently been widely discussed in the context of official statistics (Buelens et al. 2014; Daas et al., 2015).

The main goal of this paper is to assess selectivity and representativeness of IDS (web-scraped data) for the real estate market (REM) in Poland. The objective will be achieved in three steps: (1) a description of web-scraping as a method of collecting statistical data and an account of co-operation between the University and private companies that held these data; (2) a definition of weak and strong selectivity in the case of aggregated data available; (3) measurement of representativeness defined by comparing trends estimated from the IDS and official statistics. The choice of the real estate market for illustration was motivated by (i) the importance of the REM for the economy, (ii) insufficient information coverage of the REA in Poland and (iii) the importance of the Internet as a source of information on the REA. The study was conducted on the basis of domain (city level) quarterly data for the period between 2012 and 2014 from 5 IDS about the secondary REM. Research conducted by the National Bank of Poland and the Central Statistical Office in Poznań, as well as register data on transactions were used as references. All calculations were made in **R** statistical package (R Core Team, 2015).

Several web-scraping algorithms were developed to enable the continuous monitoring of the real estate market in Poland. These algorithms automatically collected data concerning 16 biggest cities in Poland (capitals of Voivodships, NUTS2 level) directly from selected web-portals. In addition, thanks to co-operation established earlier with three companies owning real estate market web-portals, it was possible to obtain historical and collect current data for the research.

Since access to individual-level data was not possible for all periods, the analysis of *selection bias* at the domain level was conducted by means of a linear mixed model (Zhang, 2012). Based on estimated bias, the author will propose a definition of weak and strong selectivity. Weak selectivity is defined to be present when the data source effect is observed, whereas strong selectivity occurs when there is interaction between domain and a data source. Several cases will be discussed in detail.

Finally, in order to assess representativeness for key variables, such as price for m^2, number of rooms and floor area of flats, a trend approach is proposed. Due to possible bias of IDS-based point estimates, a direct comparison of such levels can be misleading. Therefore, an alternative approach is put forward. It is based on the comparison of estimated trends in IDS and official data to assess whether the changes in time are coherent with official statistics.

The paper is the part of the research is supported by the National Science Centre, Poland, Preludium 7 grant no. 2014/13/N/HS4/02999.

References

Buelens, B., Daas, P., Burger, J., Puts, M., van den Brakel, J. (2014) Selectivity of Big Data. Discussion paper 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands.
Daas, P.J.H., Puts, M.J., Buelens, B., van den Hurk, P.A.M. (2015) Big Data and Official Statistics. Journal of Official Statistics 31 (2), in press.
R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Zhang, L.-C. (2012), On the accuracy of register-based census employment statistics, Q2012, Athens.

SAMPLE SURVEY OF FAMILY TO IDENTIFY THE INTENTIONS ON HAVING CHILDREN

Iryna Andras
Institute of Sociology of NAS, Belarus, androsita@tut.by

Andrei Piliutsik
Institute of Economics of NAS, Belarus, pilutic@gmail.com

Anastacia Bobrova
Institute of Economics of NAS, Belarus, nastassiabobrova@mail.ru

According to demographic forecast to 2030, the population of the city of Minsk will cease to grow. Even faster is to decrease the proportion of people of working age. At the same time the number of children ages 0-15 will change slightly, but the number of people of retirement age will increase greatly. To study the social and economic factors influencing the reproductive attitudes of families in Belarus, conducted the special sociological survey.

Date - July 2014.

The object of study - reproductive attitudes of families of Minsk. Item - social factors that affect the reproductive attitudes of families. The main hypothesis of the study - for additional motivation for the birth of the second and third child (except taken measures to support the local authorities) are strongly influenced by various types of socio-economic factors that shape life priorities and strategy for socio-demographic behavior of families.

For survey used a probability sample combination: multi-stage area sample using stratified sampling method. The first stage was selected district of Minsk (the survey was conducted in all areas of representation without requiring this degree of selection), the second - the family - the third respondent. Differentiating features fertile age.

The basis of the sample required for the characterization of the population, accounted for statistics in Minsk the National Statistical Committee of the Republic of Belarus. To determine the type and amount of sample were used statistics on the number of households and the number of residents of Minsk in the reproductive age groups. To improve the reliability of the sample of sociological research the sample size was doubled and is defined in 800 observation. This reduced the confidence interval to 0.025 or 2.5% statistically significant errors and increase reliability by combining sample stratification with probability sampling. Increased reliability sampling allows sampling error of up to 3%.

The results. The survey 56.5% indicated that they have children, and 43.5% that have no children. At the time of the survey among the answer is positive for the presence of children they have one child was 44.8%, two-child - 36.0%, three or more children - 19.2%.

The question about ideal family immediately helped identify the ideal number of children in the family by comparing the answer with information on the planned and desired number of children and to check whether respondents distinguish these concepts. In the "ideal" family of concepts 84.7% of the respondents must be sure more than one child, with one in four believes that in an ideal family should have more than three children.

In fact, the most common model for family in Minsk is a two-child family.

The main factors preventing the birth of children classified such as: material (44.9%), psychological (38.3%) and physiological (24.6%).

Determining the socio-economic conditions of family planning in the estimates surveyed Minsk residents are: material wealth (70.2%), housing, financial ability to provide a quality level of education for their children. The main permissive social factor called matrimonial status - 50.5%, and the improvement and quality of care. Important social and psychological conditions are the feeling of confidence in the future (41.7%), a favorable psychological family climate.

The main conclusion of the study is the reproductive attitudes formed under certain material conditions may be a factor of the actual reproductive behavior.

References

Shakhotska L. et al. (2014) The study of the social factors affecting the reproductive attitudes of families and forecast analysis of indicators of demographic development in Minsk until 2030. MNIISEP, 2014. - 210 p.
Sociologist's Workbook (1983). Moscow, Science.
Churilov N. (2008). Typology and design of selective sociological survey. Kyiv.

BIG DATA: ONE APPROACH TO PROCESSING ATM DATA

Iana Bondarenko
Oles Honchar Dnipropetrovsk National University, Ukraine, iana.s.bondarenko@gmail.com

Valery Turchyn
Oles Honchar Dnipropetrovsk National University, Ukraine, vnturchyn@gmail.com

Today all banks use a wide network of automatic teller machines (ATM). Such a network creates various problems to the banks. One of the main problem is that how much cash amount should be uploaded into the ATM. Uploading large cash amount leads to an increase in loss of profit through dormant assets in ATM, uploading small cash amount leads to increase in the currency transportation and servicing costs.

Bank staff have solved the problem of uploading cash amount intuitively, empirically by observing the work of each ATM on-line. But this approach is not effective, because it requires additional costs and depends on many other objectively random factors, such as problems with transport (traffic jams), weather conditions and so on. The development of effective method is needed to organize the process of uploading cash in ATM correctly and most convenient in the interests of the bank and the clients.

The work of an ATM should be organized so that *standard service quality* is implemented. Today it is accepted that at least one note should be in an ATM before the time arrival of encashment (in other words, cash should always be in ATM), and the maximum unloading cash amount must not exceed 10% of the uploading cash amount. Costs for support such standard service quality are very high. In this paper a probability model of the work of an ATM is presented to forecast uploading cash amount. Standard service quality is defined as probability of rejection for client in withdrawing cash.

Problem formulation. Let S_{up} is uploading cash amount, Δ - critical cash balance at which it is necessary to appoint ATM encashment. What should be S_{up} and Δ in order to costs for ATM encashment were minimal for a given standard service quality?

Standard service quality. Let S is the value of accumulated cash amount which were withdrawn till the moment t. Denote $S_k=S_{up}-\Delta$, where S_k is the value of accumulated cash amount at which it is necessary to appoint ATM encashment. Let $\tau(S_k)$ is moment when the accumulated amount S exceeded the S_k for the first time. For given constant α

$$P\{S-S_k \geq \Delta\} \leq \alpha$$

on the segment $[\tau(S_k), \tau(S_k)+24]$.
It is means that if we select α, for example, is *0.01*, then *99%* of clients who applied for cash, get it at this ATM during 24 hours after the moment $\tau(S_k)$. The higher standard service quality requires the greater cash balance in the ATM before 24 hours till encashment. We regulate quality of service using the level Δ.

Distribution of accumulated cash amount S which were withdrawn and distribution of the moment $\tau(S_k)$ when the accumulated amount S exceeded the S_k for the first time were identified. Optimal uploading cash amount S_{up} and optimal cash balance Δ before 24 hours till encashment, which provides *99%* standard service quality were determined.

References

Simutis R. Cash demand forecasting for ATM using neural networks and support vector regression algorithms/ Simutis R., Dilijonas D., Bastina L. // EurOPT' 2008, May 20-23, 2008, Neringa, Lithuania: selected papers. Vilnius. p. 416–421.

ON DIFFERENT POINTS OF VIEW TO THE STUDY OF SURVEY STATISTICS

Nataļja Budkina
Riga Technical University, University of Latvia, natalja.budkina@rtu.lv

Mārtiņš Liberts
University of Latvia, pm90015@lu.lv

The report deals with teaching survey statistics at the University of Latvia. There are two courses devoted to survey sampling. The first one "Survey Sampling" is given in the professional bachelor's programme "Mathematical Statistics". The second one "Survey Statistics" is included in the master's programme „Mathematics".

The special course in Survey Sampling was introduced at the Department of Mathematics in 1996 by Dr. Math. J. Lapiņš. It was intended for the fourth-fifth year students of the programme of the professional studies in Mathematical Statistics. For the reaccreditations of this programme in 2013 the course was reworked by Dr. Math. N. Budkina. Some changes should have been made in the programme of this course taking into account the development of Survey Sampling Theory and Methodology during the last years.

In 2014 the new course "Survey Statistics" was suggested to the students of the master's programme by Dr. Math. M. Liberts. The main aim of this course is to introduce students with actual problems of survey statistics – the usage of auxiliary information and non-response. "Survey Statistics" could be considered as natural continuation of „Survey Sampling". This talk presents a short description of these courses, their theoretical parts and practical works and different activities connected with the theme of survey.

There are different points of view regarding the study of methods of survey between the students and teachers at the Department of Mathematics at the University of Latvia. Some of them think that these courses are very easy, the other think that they are very difficult. On the one hand, the subject of these courses is popular among the students (the courses are not mandatory) and interest in this theme has increased during the last two years. On the other hand, there are students who pass the exam badly or with great delay. The report introduces some opinions on the study of survey statistics which were obtained from the students' questionnaires.

ASPECTS OF SAMPLING USAGE FOR RARE POPULATIONS FOR LABOR MIGRATION MEASURING IN UKRAINE

Mariia Chebanova
Taras Shevchenko National University of Kyiv, Ukraine, m_chebanova@ukr.net

Ukraine is a country with significant foreign labor migration – by some estimates it's one of the largest donors of labor force in Europe. In order to measure the scale of labor migration in Ukraine and to estimate its impacts several special state sample surveys of labor migration have been conducted over the last 10 years with the assistance of international organizations and national funds. Thus, in 2008 the first nation-wide sample survey of population on labor migration was conducted. Herewith the size of stratified multistage probability sample was about 25, 4 thousand households. In 2012 another survey was conducted within EU-funded project "Effective Governance of Labour Migration and its Skills Dimensions" and implemented by the ILO and IOM (International Organization for Migration). The last survey was conducted in 2014 – 2015 with the sample design as in the 2008 study and sample size of about 25,2 thousand households (within IOM's project "Research and Policy Dialogue Initiative on Migration and Remittances in Ukraine", funded by the Government of Canada).

A characteristic feature of the results of these surveys is a certain underestimation of the migration scale in comparison with existing expert estimates and those based on alternative sources of information. For example, Pozniak (2012) argues the results of the nationwide sample survey of labor migrations (2008) estimating the total number of Ukrainian labor migrants at 2.1 mln (vs 1,5 mln in the official report). Also the numbers presented in Labor Migration Survey (2012) raise a few questions regarding the sampling procedure (this study suggests that only 1.2 mln people between 15 and 70 were working or searching for employment abroad between January 2010 and June 2012).

Some experts suggest that given the characteristics of the labor migrants distribution, the use of standard sampling techniques for the investigation of this phenomenon is not effective. In this paper an attempt is made to estimate the probable effect of the use of rare populations sampling methods.

On the basis of synthetic population the differences of estimates of the labor migration scale with different types of distribution of migrants in the population were analyzed. In the report it was shown that in the case of rare and unevenly distributed population the adaptive sampling design may be more effective compared to conventional sampling design. With regard to the problem stated above, the technique itself as well as its advantages and limitations are discussed as a part of the attempt to obtain more precise estimates of Ukrainian labor migrants.

References

[1] Report on the Methodology, Organization and Results of a Modular Sample Survey on Labour Migration in Ukraine / International Labour Organization, Decent Work Technical Support Team and Country Office for Central and Eastern Europe (DWT/CO-Budapest). – Budapest : ILO, 2013. – 96 p.
[2] Verma V. Sampling elusive populations: Applications to studies of child labour / V. Verma ; International Labour Organization, International Programme on the Elimination of Child Labour (IPEC), Department of Statistics. – Geneva : ILO, 2013. – xix + 821 p.

RELATIONSHIP BETWEEN BALANCED SAMPLING AND CALIBRATED ESTIMATOR

Ieva Dirdaitė
Vilnius Gediminas Technical University, Lithuania, dirdaite.ieva@gmail.com

The aim of the presentation is to compare the results of estimation of a finite population total in the case of two sampling strategies: balanced sampling ([3]) of clusters with the Horvitz-Thompson estimator of total ([2]) and simple random sample of clusters with the calibrated estimator of total ([1]). The same auxiliary variables are used for sample selection in balanced sampling design and for the calibration estimator in the second strategy. The comparison is carried out by simulation. Sample data of a real Labour force survey of Statistics Lithuania, size 20,000, is used as a finite population for simulation study. Different sample sizes are used for simulation. Estimates of totals, estimates of variance estimators and mean squared errors, estimates of relative mean squared errors are compared. Conclusions and recommendations for practical surveys are drawn.

References

1. Deville J.-C., Särndal C.-E., Calibrated Estimators in Survey Sampling, *Journal of the American Statistical Association*, 1992, 87, p. 376 - 382.
2. Särndal C.-E., Swensson B., Wretman J., *Model Assisted Survey Sampling*, New York: Springer-Verlag, 1992.
3. Tillé Y., *Sampling Algorithms*, New York: Springer, 2006.

ACCURACY OF IMPUTATION: A CASE STUDY ON THE FINNISH LABOUR FORCE SURVEY

Kari Djerf, Atte Lintilä, Riku Salonen, Ari Veijanen
Statistics Finland
Contact: firstname.surname@stat.fi

The Finnish Labour Force Survey does not routinely utilize imputation methods in standard statistics production. In January 2015 the electronic interview questionnaire contained a programming error which resulted in a missingness: for a group of respondents, one crucial question was completely passed by. It is one of those to determine whether the respondent is unemployed according to Eurostat definition or not. After finding the mistake all cases involved were re-interviewed. However, about one third of the respondents could not be re-contacted and the information was then imputed by using both traditional cold deck and nearest neighbour donor imputation methodology. We will present the results and evaluate the accuracy of imputations. The evaluation will be carried out by comparing the imputed values with true values from the re-interviews when available. The labour force status of those cases having subsequent interview round three months later were also compared with the imputed values.

INCOME INEQUALITY MEASURES FOR BALTIC AND NORDIC COUNTRIES

Kristina Galiautdinovaitė
Vilnius Gediminas Technical University, Lithuania, kristina.galiautdinovaite@gmail.com

Income inequality measures are estimated by official statistics of many countries. Gini coefficient is one of such measures. The purpose of this study is to estimate Gini coefficient for Baltic and Nordic countries using data of European Social Survey 2012. Specific sampling design of each country will be taken into account. The results obtained will be compared with the officially published statistics.

References

1. *ESS Cumulative Data*, 2015. Retrieved from http://www.europeansocialsurvey.org/.
2. European Commission, Eurostat. *Theoretical study of the Gini index*. EU-SILC 131-A/04, June 2004.
3. World Bank Institute, *Introduction to Poverty Analysis*, 2005, p. 95-105.

ARE WE WITNESSING THE END OF RANDOM SAMPLING IN SURVEYS?

Dan Hedlin
Department of Statistics, Stockholm University, dan.hedlin@stat.su.se

The method of random sampling has an extremely strong position in official statistics. In the past decades, high quality surveys based on random sampling have become more difficult and expensive to conduct due to a hardening survey climate. The issue whether we should prefer random sampling to non-random methods pushes itself to the forefront of the agenda.

Recent research on 'balancing' (Särndal & Lundquist 2014) or 'representativeness' in sample surveys seems promising, although this is still ongoing research. I will argue that this may pull us away from random sampling towards obtaining a 'good' set of data that will allow for reliable inference, no matter whether that good data set has been obtained through a random sampling mechanism or not.

Reference

Särndal, C.-E. and Lundquist, P. (2014). Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation. Journal of Survey Statistics and Methodology, 1-27.

SYSTEMATIC HANDLING OF MISSING DATA IN COMPLEX STUDY DESIGNS - EXPERIENCES FROM THE HEALTH 2000 AND 2011 SURVEYS

Tommi Härkänen
National Institute for Health and Welfare, Finland, tommi.harkanen@thl.fi

Juha Karvanen
University of Jyvskyl, Finland, juha.karvanen@jyu.fi

Hanna Tolonen
National Institute for Health and Welfare, Finland, hanna.tolonen@thl.fi

Risto Lehtonen
University of Helsinki, Finland, risto.lehtonen@helsinki.fi

Kari Djerf
Statistics Finland, Finland, kari.djerf@stat.fi

Teppo Juntunen
National Institute for Health and Welfare, Finland, teppo.juntunen@thl.fi

Seppo Koskinen
National Institute for Health and Welfare, Finland, seppo.koskinen@thl.fi

We present a systematic approach to the practical and comprehensive handling of missing data motivated by our experiences of analyzing longitudinal survey data. We consider the Health 2000 and 2011 Surveys (BRIF8901) where increased non-response and non-participation from 2000 to 2011 was a major issue. The model assumptions involved in the complex sampling design, repeated measurements design, non-participation mechanisms and associations are presented graphically using methodology previously defined as a causal model with design i.e. a functional causal model extended with the study design. This tool forces the statistician to make the study design and the missing data mechanism explicit. Using the systematic approach, the sampling probabilities and the participation probabilities can be considered separately. This is beneficial when the performance of missing data methods are to be compared. Using data from Health 2000 and 2011 Surveys and from national registries, it was found that multiple imputation removed almost all differences between full sample and estimated prevalences. The inverse probability weighting removed more than half and the doubly robust method 60% of the differences. These findings are encouraging since decreasing participation rates are a major problem in population surveys worldwide.

SOME APPROACHES IN ANALYZING THE DATA WITH EXCESS OF ZEROS

Tetiana Ianevych
Taras Shevchenko National University of Kyiv, Ukraine, yata452@univ.kiev.ua

I examine the different models and corresponding LM estimators for data containing many zero values and analyze their usefulness for designing sample survey of capital expenditure in Ukraine.

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

In order to avoid underestimation of the capital expenditure value in the quarterly surveys the probabilistic sampling should be implemented into the investigation. Since the large and medium-sized enterprises are more valuable they are all observed every time. We survey by the means of the sample only the small and new enterprises. Thereby we focus on the investigation of small enterprises and perceive them as population.

I have made several attempts to incorporate into the designing and estimation process the additional information, e.g. the data on expenditure from the previous surveys, revenue, number of employee, etc. There were utilized the ordinary regression estimator and different robust estimators in order to obtain more accurate estimates not only for the estimate on the country level but also for different domain estimates. This led us to the problem of dealing with many zero values.

There are a number of econometric approaches to dealing with the problem of zeros. These approaches differ depending on the type of the data. If data is countable than statisticians often used such model as zero inflated Poisson or similar. Capital expenditure is not a count data, so we need to use the models that deal with semi-continuous data. It means that it has a continuous distribution except for a probability mass at 0. Good review paper devoted to all this models is a paper by Min&Agresti (2002).

So, we decided to compare the following estimators for the estimation of the capital expenditure in 2010: Horwitz-Tompson; GREG with and without log-transformation of independent variable which is capital expenditure in 2009; regression estimator based on Tobit model with capital expenditure in 2009 as a regressor; and a regression based on Heckit model capital expenditure in 2009 as a regressor in the outcome equation and identification variable on whether the enterprise had expenditure last year as a regressor for selection equation. We have made 1000 Monte Carlo simulation for all these estimators and calculated the ARB and RRMSE. The results of this simulation study are presented in the Table 1.

Table 1: Comparison of different estimators

	HT	GREG	Log-transform + GREG	Tobit	Heckit	Log-transfor + Heckit
ARB, %	0.58	10.18	0.72	24.18	4.09	0.41
RRMSE, %	41.70	36.02	37.77	51.18	40.88	41.41

As we can see, usage of GREG estimator leads to biased but better results with regards to the accuracy. The usage of the Tobit and Heckit-based estimators fell short of expectations but still can be improved by changing or incorporating more independent variables. And the main useful thing is that all theses estimators except HT can be used for the construction of small area estimators.

References

Min, Yo., Agresti, (2002) A., Modeling nonnegative data with clumping at zero: A survey. JIRSS, Vol. 1, Nos. 1-2, 7-33.

IMPROVING EFFICIENCY OF THE SAMPLE DESIGN IN THE FINNISH HORTICULTURAL SURVEY

Anna-Kaisa Jaakkonen
Natural Resources Institute Finland, Finland, anna-kaisa.jaakkonen@luke.fi

Mika Kuoppa-aho
Natural Resources Institute Finland, Finland, mika.kuoppa-aho@luke.fi

Johanna Laiho-Kauranne
Natural Resources Institute Finland, Finland, johanna.laiho-kauranne@luke.fi

The Finnish horticultural survey has been conducted annually as a total survey with a threshold on standard economic output (SO) of the horticultural enterprises. The standard output of an agricultural product (SO), is defined as the average monetary value of the agricultural output at farm-gate price, in euro per hectare or per head of livestock. Traditionally the threshold has been relatively low in the horticultural survey in comparison to the average wages of the household for example. Therefore, the impact of the small horticultural enterprises for the final survey estimates is studied including their impact on the survey costs and on the quality of the survey data.

In this paper, we investigate the impact of increasing the threshold both on the quality and coverage of the final estimates as well as the impacts on the survey cost component. With the sensitivity analysis using previous survey data, we can present the detection of the optimal threshold on the standard economic output to balance the survey costs and the quality criteria of the survey defined in the EU regulation survey for permanent crops. We also present the method of deriving and monitoring the development of the standard economic output for the horticultural enterprises in the sampling frame.

The horticultural survey uses both the auxiliary data from statistical farm and horticultural register and collects directly information about the production of horticultural products, as well as on the production and the use of energy at the horticultural enterprises. To reduce the survey costs, and to improve the quality of the survey data, the sampling design of the survey is reviewed in detail. The data collection of the horticultural survey uses mixed-mode approach; using register data, web-survey, telephone interviews and accepts also paper questionnaires. With the increase of the thresholds we can also analyse the increase in the web-survey response rate. Thus it is expected that larger horticultural enterprises tend to respond through the web-survey more likely; while those who are interviewed tend to be on average smaller enterprises. Therefore, we will also present the impact of the efficient sample design on the expected improvement on the timeliness of the survey data.

References

Eurostat (2015): *Statistics explained. Glossary:Standard output (SO).*
http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Standard_output_(SO)
Natural Resources Institute Finland (2015): *Horticultural Statistics 2014 (final statistics) and Energy Consumption in Greenhouse Enterprises.*
http://stat.luke.fi/en/horticultural-statistics
Dillman D.A., Smyth J.D., Christian, L. M. (2009): *Internet, mail and mixed-mode surveys. The Tailored Design Method.* 35d ed. Wiley.
Snijkers, G., Haraldsen, G., Jones, J., Willmack, D. K. (2013): *Designing and conducting business surveys.* Wiley series in Survey Methodology.

GENETIC MODELLING IMPUTATION APPROACH FOR QUANTITATIVE MISSING DATA PROBLEM

Aydın KARAKOCA
Necmettin Erbakan University, Turkey, akarakoca@konya.edu.tr

Alper SİNAN
Sinop University, Turkey, alpsin@sinop.edu.tr

Researchers always encounters with missing data problem. Since 1980s, many imputation methods proposed for example Little(1976), Little and Rubin(2002) and Schafer(2010). Various imputation methods and different analysis techniques are developed for dealing with the missing data. in literature as mean imputation, regression imputation etc.

In this study we proposed genetic modelling imputation approach in order to increase the performance of regression imputation methods. Genetic modelling imputation (GMI) method aims to find the best functional form generated by genetic algorithm. Performance and imputation accuracy of GMI ,regression imputation and mean imputation methods compared for four different missing ratios 1%,%5,10% and 15% by simulation and a real data application is given.

References

Little, R. J. A., and Rubin, D. B., 2002, Statistical analysis with missing data, John Wiley&Sons.
Little, R. J. A ,1976, Inference about Means from incomplete Multivariate data, Biometrika,63,3,593-604.
Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. Machine learning, 3(2), 95-99.
Schafer,J.L.,2010,Analysis of Incomplete Multivariate Data,CRC press.

SURVEY DATA AND BAYESIAN ANALYSIS: A COST-EFFICIENT WAY TO ESTIMATE CUSTOMER EQUITY

Juha Karvanen
University of Jyvaskyla, Finland, juha.t.karvanen@jyu.fi

Ari Rantanen
Sanoma Media Finland, Finland

Lasse Luoma
Tietoykkönen Oy, Finland

We present a Bayesian framework for estimating the customer lifetime value (CLV) and the customer equity (CE) based on the purchasing behavior deducible from the market surveys on customer purchasing behavior. The proposed framework systematically addresses the challenges faced when the future value of customers is estimated based on survey data. The scarcity of the survey data and the sampling variance are countered by utilizing the prior information and quantifying the uncertainty of the CE and CLV estimates by posterior distributions. Furthermore, information on the purchase behavior of the customers of competitors available in the survey data is integrated to the framework. The introduced approach is directly applicable in the domains where a customer relationship can be thought to be monogamous. As an example on the use of the framework, we analyze a consumer survey on mobile phones carried out in Finland in February 2013. The survey data contains consumer given information on the current and previous brand of the phone and the times of the last two purchases.

References

J. Karvanen, A. Rantanen, L. Luoma (2014). Survey data and Bayesian analysis: a cost-efficient way to estimate customer equity. *Quantitative Marketing and Economics*, 12(3), 305–329, DOI:10.1007/s11129-014-9148-4, arXiv:1304.5380.

REDUCTION OF RESPONSE BURDEN BY REPLACING SURVEY QUESTIONS WITH REGISTER DATA: CASES OF CROP ROTATION, NON-REGULAR NON-FAMILY LABOUR, AND NUMBER OF ANIMALS

Esa Katajamäki
Natural Resources Institute Finland, Finland, esa.katajamaki@luke.fi

Pasi Mattila
Natural Resources Institute Finland, Finland, pasi.mattila@luke.fi

Johanna Laiho-Kauranne
Natural Resources Institute Finland, Finland, johanna.laiho-kauranne@luke.fi

In ESS statistics Eurostat supports the use of administrative data. Natural Resources Institute Finland (Luke) has a long experience of using administrative registers as a source for statistics. However when the administrative data develops, new possibilities arises. Our objective is to replace survey data in forthcoming farm surveys with register data, and to develop required survey estimation procedures. We examine the usability of the agricultural registers data to reduce the response burden and reducing the information collected directly from the farms:
1) Crop rotation (FSS variable "Share of arable land included in crop rotation"), source register: IACS parcel data from the Finnish Agency for Rural Affairs.
2) Farm relief workers' amount of work (part of the FSS variable "non-family labour employed on a non-regular basis"), source register: The Finnish Farmers' Social Insurance Institution.
3) Number of pigs, sheep and goats, source registers: Pig register and Sheep and Goat register maintained by the Finnish Food Safety Authority.

The general objectives in our study are:
1) To investigate possibilities for a broader analysis of crop rotation based on the IACS parcel data icluding the new geospatial parcel data obtained from the farmers through farm subsidy administration from the year 2015 on.
2) To pilot the use and examine the quality of the individual level register information that can be linked with identificators to farm level from a register other than IACS as a source of FSS data.
3) To evaluate the quality and feasibility of Pig register and Sheep and Goat register as sources of data for animal statistics and FSS.

Common advantages of the use of registers are the almost total coverage of farms and the avoidance of misinterpretations by farmers when answering the questionnaires which is a significant factor in the case of the crop rotation variable, for example.

References

Lehtonen, R. & Pahkinen (2003*) Practical Methods for Design and Analysis of Complex Surveys*, 2nd Edition. Wiley.
Eurostat (2003) *Methodology of animal statistics. A study of the methodology applied by the Member States of the European Union and candidate countries to livestock surveys, slaughter statistics, production forecasts (gross domestic production), external trade statistics and the latest developments in the field of poultry statistics.* European Comission. Theme 5 Agriculture and fisheries.
Gebremedhin, B. & Schwab, G. (1998). *The Economic Importance Of Crop Rotation Systems: Evidence From The Literature*. Staff Paper No. 98-13. Department of Agricultural Economics Michigan State University.

MODEL-BASED OPTIMAL SAMPLE ALLOCATION FOR PLANNED AREAS USING EBLUP ESTIMATION

Mauno Keto
Mikkeli University of Applied Sciences, Finland, mauno.keto@mamk.fi

Erkki Pahkinen
University of Jyväskylä, Finland, erkki.pahkinen@jyu.fi

This paper studies sample allocations used in surveys when it is a question of model-based area estimation and an area coincides with a stratum of stratified sampling. Although model-assisted and model-based estimations are common in the production of area statistics, utilization of the used model and estimation method are not included in sample area allocation solutions. This is the reason why two model-based allocation methods, which deploy a model and an estimator, have been developed here. The first one, g1-allocation, is based on a measure of homogeneity within areas measured of an auxiliary variable, and Sim-allocation is based on sample sizes obtained from medians of area-specific simulation results. To compare efficiency of new developed methods five allocation methods are picked up from literature. Empirical comparisons of performances of different area allocations are based on EBLUP estimation results obtained from simulated samples.

Key words: optimal area sample size, criteria, auxiliary information, homogeneity measure.

References

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician* **42** 174-177.
Choudhry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for effective domain estimation. *Survey Methodology* **38**, 23-29.
Costa, A., Satorra, A. and Ventura, E. (2004). Improving both domain and total area estimation by composition. *SORT* **28** (1) 69–86.
Keto, M. and Pahkinen, E. (2014). On sample allocation for efficient small area estimation. *Book of Abstracts*. SAE 2014 page 50. Poland: Poznan University of Economics.
Longford, N. T. (2012). Allocating a Limited Budget to Small Areas. *Journal of the Indian Society of Agricultural Statistics* **66** 31–41.
Molefe, W. and Clark, R.G. (2013). Model-assisted optimal allocation for planned domains using composite estimation. University of Wollongong, *Working Paper* 19-13, 1-27, http://ro.uow.edu.au/cssmwp/109.

TRENDS IN CERVICAL CANCER INCIDENCE IN LATVIJA IN 1983-2013

Una Kojalo
Riga Stradins university, Statistics unit, Una.Kojalo@rsu.lv

Ģirts Briģis
Riga Stradins university, Public Health and Epidemiology department, Latvia, Girts.Brigis@rsu.lv

Cervical cancer is the third most common cancer form in the world for woman and the fourth most common cause of woman death [1]. Cervical cancer incidence is closely linked to the quality of health care system in the country. Various developed countries reached a significant reduction in the cervical cancer incidence with well organised cancer screening programmes [2]. Cervical cancer is the sixth most common cancer site for Latvian women after breast, skin, colorectal, uterine and ovarian cancer [3].

The aim of the study is to calculate age standardized cervical cancer incidence rates and to show the incidence time trends as indicators of changes in health care system in Latvia during past thirty years.

The study included data obtained from the population-based Latvian cancer registry. The sample included 5890 women with diagnosed and histologically confirmed cervical cancer in the period from 1983 to 2013. Age-standardized rates were calculated by direct standardization method using world standard population [4],[5]. Incidence changes were detected with join point regression method using the National Cancer Institute program Joinpoint Software 4.1.0 [6]. Connection point regression is a technique that is used to analyse changes over time. With calculations are defined time periods when rates change linearly, as well as the time points at which these periods are changing. Each calculation for the period is fixing annual percentage changes (APC) and its 95% confidence interval. Data processed using MS Excel 2010 and IBM SPSS 20.0 programs.

Cervical cancer incidence trends changes twice, in 1990 (95% CI: 1988 – 1993), and 1993 (95% CI: 1992 – 1998) divided incidence trend into 3 periods: in the period 1983 to 1990 incidence was initially decreasing, in 1990 - 1993 increasing dramatically and in 1993 - 2013 continues increasing with an annual percentage of 3,8% (95% CI: 3,1 – 4,6). Early cancer stage group have two major time periods: 1983 - 1992, an annual percentage decrease is 3.2% (95% CI: 0,8 - 5,5) per year; and 1992-2013 APC started increase of 2.7% (95% CI: 2,0 – 3,4) per year. Advanced cancer forms trends changed twice: until 1989 trends did not changed significantly, in the period till 1993 has increased dramatically, with 51.6% (95% CI: 24,6 – 84,4); the third period, from 1993 to the 2013th, APC continues to grow by 3.8% (95% CI: 2, 7 – 4,9). Cervical cancer incidence trends shows changes during different economic and political periods in Latvia [7], however trends does not show any changes since organised screening programme started in Latvia in 2009.

References

1. Jemal, A., et al., *Global cancer statistics*. CA Cancer J Clin, 2011. **61**(2): p. 69-90.
2. Vaccarella, S., et al., *50 years of screening in the Nordic countries: quantifying the effects on cervical cancer incidence*. Br J Cancer, 2014.
3. SPKC, *Onkoloģija. Statistikas dati par pacientu skaitu sadalījumā pa reģioniem, lokalizācijas veidiem, dzimuma un vecuma grupām no 2009.gada līdz 2013.gadam.*, 2014, Slimību profilakses un kontroles centrs. p. 1-19.
4. Ahmad, O.B., et al. *Age standardization of rates: a new WHO standard*. 2001 sk. 30.06.2014.; Available from: http://www.who.int/healthinfo/paper31.pdf.
5. Bray, F., *Age-standardization.* Cancer incidence in five continents, 2002. **8**: p. 87-89.
6. NCI. *Joinpoint Regression Program*. 2014 sk. 30.06.2014.; Available from: http://surveillance.cancer.gov/joinpoint/.
7. Viberga, I. and M. Poljak, *Cervical cancer screening in Latvia: a brief history and recent improvements (2009-2011).* Acta Dermatovenerol Alp Pannonica Adriat, 2013. **22**(1): p. 27-30.

CORRECTING FOR NON-IGNORABLE MISSINGNESS IN HEALTH INDICATOR TRENDS

Juho Kopra
University of Jyvaskyla, Finland, juho.j.kopra@jyu.fi

Tommi Härkänen
National Institute for Health and Welfare, Finland, tommi.harkanen@thl.fi

Hanna Tolonen
National Institute for Health and Welfare, Finland, hanna.tolonen@thl.fi

Juha Karvanen
University of Jyvaskyla, Finland, juha.t.karvanen@jyu.fi

Data missing not at random (MNAR) are a major challenge in survey sampling. We propose an approach based on registry data to deal with non-ignorable missingness in health examination surveys. Our approach relies on follow-up data available from administrative registers several years after the survey. For illustration, we use data on smoking prevalence in the Finnish FINRISK study. The data consist of survey information for the participants and including missingness indicators, register-based background information and register-based time-to-disease survival data for the full sample. The parameters of missingness mechanism are estimable with these data although the original survey data are MNAR. The underlying data generation process is modelled by a Bayesian model. Our results indicate that the estimated smoking prevalence rates in Finland may be significantly affected by missing data.

SMALL AREA ESTIMATION FOR A STUDY VARIABLE HAVING MANY ZERO VALUES

Danutė Krapavickaitė
Vilnius Gediminas Technical University, Lithuania, danute.krapavickaite@vgtu.lt

Any real sample survey is carried out in order to estimate parameters not only for finite population, but also for many domains. Of course, these domains may be taken into account when constructing a stratified sample design. If it is not possible for some reasons, then it may occur that the sample size for design-based estimator in some domains may be too small to obtain sufficiently accurate estimate, or may be no sampled elements in some domains at all. The usage of auxiliary information at the estimation stage may improve the situation. The generalised regression estimator is one of the possibilities. The model-based estimator is another possibility. Estimation methods for the areas having small sample size (small areas, [4]) is a very popular topic of survey sampling nowadays. We will talk about model-based small area estimation for a study variable having many zero values ([2,3]). Bayesian inference will be used for this ([1,5]).

References

1. Geweke J. (2003), *Contemporary Bayesian Econometrics and Statistics*, University of Iowa.
2. Greene W. H. (2003), Econometric Analysis. Prentice Hall, Upper Saddle River.
3. Krapavickaitė D. (2011), Some models for estimation of total of a study variable having many zero values, *Lith. Mathemm. J.*, 51(3), p. 370-384.
4. Rao J. N. K. (2003), *Small Area Estimation*, Hoboken, John Wiley and Sons.
5. Statisticat, LLC (2014). *LaplacesDemon: Complete Environment for Bayesian Inference*. R package version 14.04.05, URL http://www.bayesian-inference.com/software.

STATISTICAL ESTIMATION AND ANALYSIS OF FOREIGN TRADE IN HEALTH SERVICES OF THE REPUBLIC OF BELARUS

Anna Larchenko
Ministry of Foreign Affairs of the Republic of Belarus, Belarus, annalarchenko@gmail.com

According to the methodology of payment balance main types of international services in the Republic of Belarus are: transportation services, travel (tourism services), communication services, construction services, insurance services, financial services, computer and information services, government services, other business services.

According to the methodology of official statistics exports of services – provision of different types of services by the residents of the Republic of Belarus to non-residents. Imports of services – receiving by the residents of the Republic of Belarus various services from non-residents. Service cost is performed according to the delivery contract or according to any document confirming the provision of services.

Official statistical data are formed on the base of monthly statistical reporting "Report on the exports and imports of services". One of the respondents is the Ministry of Foreign Affairs of the Republic of Belarus (aggregated primary statistical data on the diplomatic missions of the Republic of Belarus abroad).

In accordance with the official statistical methodology health services are included into the group of government services. Health services include services for diagnosis and treatment of diseases, including health insurance; consulting and other services in the health sector; services provided by medical personnel abroad, including the services provided in absentia.

To collect the information of how people in reproductive age (men of age 15-59 and women of 15-49) evaluate the state of their health and the quality of health services in Belarus mini-survey of Reproductive Health has been carried out by the author in Minsk.

The object of the survey is women aged 15-49 and men aged 15-59, living in Minsk. According to the author's calculations sample size in Minsk was 1010 persons (603 females and 407 males) [1]. Separate questionnaires for men and women at reproductive age have been used as survey tools. To collect data the one-stage quasi-random sample has been used [1, 2]. Survey elements have been investigated directly, without additional selection steps, and each unit has been examined once.

Main indicators of Reproductive Health Survey in Minsk are:

- Socio-economic status of the respondents;
- Anthropological characteristics of the respondents;
- Respondents' reproductive health;
- Reproductive attitudes;
- Assessment of health care quality and respondents' expenses for medicines and medical services, % of revenue.

References

[1] Larchenko, A. Reproductive Health Survey: determination of sample size and design / A. Larchenko // Summer School of Baltic-Nordic-Ukranian Network on Survey Statistics, Minsk, June 13—19, 2013 / Institute of Economics of National Academy of Sciences of the Republic of Belarus; edit.: N. Bokun [etc.].— Belarus, 2013. — P. 75—81.
[2] Bokun N.Ch. Methods of Survey Sampling: Tr. and Ref. Manual / N.Ch. Bokun, T.M. Chernyshova; Minisrty of Statistics and Analysis of the Republic of Belarus, Research Institute of Minisrty of Statistics and Analysis of the Republic of Belarus. — Minsk, 1997. — 416 p.

SMALL AREA ESTIMATION BY CALIBRATION METHODS

Risto Lehtonen
University of Helsinki, Finland, risto.lehtonen@helsinki.fi

Ari Veijanen
Statistics Finland, ari.veijanen@stat.fi

There is increasing demand in the society for reliable statistics for various population subgroups such as regional areas. Examples are unemployment figures for municipalities and regional poverty indicators. If regional sample sizes are small, conventional direct estimators such as Horvitz-Thompson estimator are inaccurate. Small area methods have been developed to improve accuracy over conventional methods. Small area methods use auxiliary information from registers and statistical modelling to improve the accuracy. We discuss design-based calibration methods for the estimation of totals for population subgroups or domains (small or large). The methods include the traditional model-free or linear calibration (Deville and Särndal 1992) and model calibration (Wu and Sitter 2001) and certain more recent model calibration estimators introduced in Lehtonen and Veijanen (2012, 2015) including semi-direct model calibration methods. We introduce a new variant of model calibration called "hybrid" calibration (Lehtonen and Veijanen 2014). This method aims at combining some favorable properties of model calibration (accuracy improvement; applicability to nonlinear models) and model-free calibration (coherence property of estimates with published statistics), resembling a method proposed in Montanari and Ranalli (2009). These methods assume an access to unit-level auxiliary data. This is a realistic assumption in many "register" countries such as Finland. In the paper we compare the statistical properties (design bias and accuracy) of calibration estimators with design-based simulation experiments by using unit-level synthetic populations.

References

Deville J.-C. and Särndal C.-E. (1992). Calibration estimators in survey sampling. JASA 87, 376-382.

Lehtonen R. and Veijanen A. (2012). Small area poverty estimation by model calibration. Journal of the Indian Society of Agricultural Statistics 66, 125-133.

Lehtonen R. and Veijanen A. (2014). Small area estimation of poverty rate by model calibration and "hybrid" calibration. NORDSTAT Conference, Turku, June 2014.

Lehtonen R. and Veijanen A. (2015). Design-based methods to small area estimation and calibration approach. In: Pratesi M. (Ed.) Analysis of Poverty Data by Small Area Estimation. Chichester: Wiley. (forthcoming)

Montanari G.E. and Ranalli M.G. (2009). Multiple and ridge model calibration. Proceedings of Workshop on Calibration and Estimation in Surveys 2009. Statistics Canada.

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

AUXILIARY INFORMATION IN DATA COLLECTION AND ESTIMATION STAGE

Kaur Lumiste
University of Tartu, Estonia, kaur.lumiste@ut.ee

Large non-response has become an almost unavoidable part of every survey, leading to biased results and questionable inference. There is extensive literature on how to reduce non-response bias of estimates in the estimation stage of the survey, but corrective actions can and should be taken earlier – in planning and data collection phases. There is a growing number of research being done on *responsive survey designs* (Särndal 2011), where the goal is to get a well representative set of respondents through planning and appropriate intervention in the data collection process. In current work we aspire to a "representative" set of respondents through balance of the response set with respect to a given set of auxiliary variables – means of auxiliary variables have to be approximately the same in the sample and the response set. We assess balance with an imbalance measure discussed by Särndal (2011) using auxiliary information.

The same auxiliary variables can also be used in the estimation stage to improve our estimates, but assume that we have access to more auxiliary variables in the estimation stage than we did in the data collection stage. Is the effect of additional explanation power affected by balancing? Finding an answer to this question brings another one - should we emphasise on acquiring more auxiliary variables for the estimation stage or should we focus more on balancing the response? Which would have a larger effect on the bias and/or accuracy of the final estimates? Our goal is to determine the effects of using auxiliary information both in the data collection stage and post-weighting.

References

Särndal, C.-E. (2011). The 2010 Morris Hansen lecture dealing with survey nonresponse in data collection, in estimation. Journal of Official Statistics 27, 1–21.

COMBINING HBS AND LFS DATA FOR EDUCATION LEVEL ESTIMATION

Olha Lysa
Ptoukha Institute for Demography and Social Studies of NASU, Ukraine, Olysa@ukr.net

The last Ukrainian Census was conducted in 2001, thus important information about social and demographic characteristics of population and households is only available in data collected in sample surveys. Two state sample surveys which are provided by the Ukrainian Statistical Service are the main information sources about population condition.

The Household Budget Survey (HBS) and the Labour Force Survey (LFS) focus on different topics but are harmonized and have a number of core questions in common. We consider these aspects and their influence on the possibility of combining HBS and LFS data.

Data matching procedures are used to produce the combined dataset with social and demographic indicators: sex, age, education level, social status, household structure and so on. A simple method of weighting is not effective for combined data, so we had to consider more efficient methods such calibration.

The calibration model includes distributions of: (1) households by regions and types of area; (2) households with children by regions; (3) population by regions and types of area; (4) population by gender, age groups and region.

The estimates of education level obtained from the combined dataset have quite high accuracy, not only at the national level but for different subgroups. These results will be presented and discussed during the presentation.

To use the statistical matching methods for HBS and LFS allows for: (1) a significant increase of sample survey data quality; (2) the possibility to obtain reliably estimates of education level at the subnational and regional levels; (3) the coordination of sample survey results with external sources; (4) comparability of data.

References

[1] Rässler, S. (1998) Aspects concerning data fusion techniques, ZUMA Nachrichten Spezial 4, 317_333.
[2] D'Orazio, M., Di Zio, M., Scanu, M. (2006) Statistical Matching: Theory and Practice, Chichester: John Wiley&Sons.

FIRST RESULTS IN DETERMINING PERMANENT RESIDENCY STATUS IN REGISTER-BASED CENSUS

Ethel Maasing
Statistics Estonia, Estonia, ethel.maasing@stat.ee

The next population and housing census in Estonia at the end of 2020 is intended to be register-based. At the moment analysts and methodologists are analysing different administrative registers to determine the total population of persons, households and dwellings. This presentation will give an overview of the first results of determining the total population of persons. The work was done for the Master's thesis in mathematical statistics at the University of Tartu. The author used logistic regression analysis and other Tartu University postgraduate students used different assumptions of logistical and linear regression and discriminatory analysis.

All the data about Estonian citizens and foreigners, who have registered their address in Estonia or have got an Estonian residence permit are collected at the Population Register (PR). All persons have a unique identification code that is also used by other administrative registers in Estonia. Everybody is obliged to register their right address to PR by the law. There have been situations where people who have left Estonia do not register their leaving in PR or people who have come (back) into Estonia do not give this information to PR. It can be assumed that the people, who actually live in Estonia are represented in other administrative registers, because they are using services and receive payments.

Besides PR there were 10 administrative registers used: Estonian Education Information System; Register of Social Services and Benefits; Health Insurance Database; National Defence Obligation Register; State Pension Insurance Register; Register of persons registered as unemployed or job-seekers, and of provision of labour market services; Register of Residence and Work Permits; E-file system; Estonian Traffic Register; Register of Employment. Furthermore last census data that was held on the end of 2011 was used to determine the actual residents and non-residents groups.

The work carried out by the author using logistic regression analysis showed that it was easiest to differentiate residents from non-residents using administrative registers among 7–16-year-olds, but most difficult among men aged 23–62. Statistics Estonia published the total population number on the basis of the personalized population base after the last census. Comparison with the analysis results and the personalized population base showed that discriminant analysis gave the closest results.

Work with the determination of total population will continue using administrative registers that were left out of the analysis because of technical reasons and the control groups will be adjusted using the proposals of the thesis.

THE FISHING MANAGEMENT FEE REGISTER AND THE POPULATION REGISTER AS SAMPLING FRAMES IN A FINNISH RECREATIONAL FISHING SURVEY

Pentti Moilanen
Natural Resources Institute Finland, Finland, pentti.moilanen@luke.fi

Anssi Ahvonen
Natural Resources Institute Finland, Finland, anssi.ahvonen@luke.fi

In 2012, there were about 1.5 million recreational fishermen (28 percent of the population) in about 850,000 household-dwellings in Finland. The total catch amounted to 24.5 million kg fish and 2.5 million pieces crayfish. About 3.5 million kg fish and one million crayfish were released alive.

Traditionally the statistics on recreational fishing has been compiled every second year using sampling surveys. The sample design has been stratified sampling. The sample has been drawn from the population register maintained by the Population Register Centre. The sample size has been 6,000 household-dwelling units.

With certain angling exceptions, fishing in Finland requires a payment of the fishing management fee. When studying recreational fishing in 2014, we were able to merge the fishing management fee register to the population register. In the sampling design household-dwellings, included into the fishing management fee register, formed one stratum.

From the survey data it is possible to study by strata the structure of the household-dwellings, using of different gear types, the catch estimates etc. This might give new opportunities to develop the sampling and the data collection methods to get more precise estimates.

Keywords: recreational fishing, stratified sampling, sampling frame

ESTIMATING THE LENGTH OF WORKING CAREERS FROM THE FINNISH LABOUR FORCE SURVEY DATA

Markku Mikael Nurminen
Department of Public Health, University of Helsinki, Finland
MarkStat Consultancy, markstat.net, Helsinki, Finland
markku.nurminen@markstat.net

Background. Demographic aging is ensued by many adverse societal consequences. The concerns are encountered in all developed countries like Finland with a very fast aging population profile. Extending the time spent in employment has been proposed as one of the key drivers for adjusting to the prolonged longevity of the population, and thereby to work out the involved public health and socioeconomic problems. Yet the measurement of the length of working careers is quite a statistical chore and currently not a standard practice.

Objectives. This paper first refers to the definitions of alternative expectancy measures and analytically re-examines the practices of statistical methods employed for estimating *working-life expectancy*, i.e. the expected number of years in employment remaining in one's life at a given age. The estimation is then done jointly for the future occupation times in the states of employment, unemployment and outside the labor force. The final aim is to explain and discuss the reasons for the discrepancies between the estimates.

Data and Methods. Because of the methodological nature of this discourse, the reviewed studies were not selected systematically, rather they were included based on their relevance, currency and high scientific quality. Expressly, the focus was on the examination of the advantages and limitations of two fundamentally unlike approaches − the period life table technique (1) and a multiple regression model for the multistate cohort life table (2) − and the comparison of their estimates derived from a population-based study (3), which was designed to analyze the aggregated annual Labor Force Survey data from Finland in the years 2000-2010 produced by Statistics Finland. Further, the model-based predictions were projected for the years 2011-2015.

Results. The study (3) found a marked difference (gap) between period (the Sullivan method) and cohort (the Davis et al. method) working-life expectancies. The methodological article (4) provided cogent arguments, substantiated by empirical findings from previous studies, to evince the superior performance of the preferred multistate vector regression approach over the traditional period life table technique for measuring working-life expectancies in epidemiological research and actuarial practice.

Conclusions. The multistate modeling and estimation methodology presented makes possible an improved statistical analysis of stochastic processes in working life. The major advantage of the multistate modeling lies in its way to reconstruct the relevant elements of the longitudinal (cohort) stochastic process that generated the working-life table datasets from annual cross-sectional surveys. This method quantifies successfully the magnitude of the national demographic problem that stresses the demand for pension reforms and other sociopolitical actions.

References

1. Sullivan DF. A single index of mortality and morbidity. *Health Services and Mental Health Administration Health Reports* 1971;86:347-54.
2. Davis BA, Heathcote CR, O'Neill TJ. Estimating cohort health expectancies from cross-sectional surveys of disability. *Statistics in Medicine* 2001;20:1097-1111.
3. Nurminen M. Working-life expectancy in Finland: trends and differentials 2000-2015. A multistate regression modeling approach. Helsinki: Finnish Centre for Pensions, *Reports* 03/2012 [100 pages]. http://www.markstat.net/en/images/stories/career_length.pdf
4. Nurminen M. Measuring working-life expectancies: Multistate vector regression approach vs. prevalence-based life table method. *The Internet Journal of Epidemiology* 2014 Volume12 Number 1. https://ispub.com/IJE/12/1/14812

COMPARISON OF MISSING DATA METHODS USING REGISTER-BASED AUXILIARY DATA FOR HEALTH-RELATED SURVEY DATA OUTCOME

Oona Pentala, Tommi Härkänen and Risto Kaikkonen
National Institute for Health and Welfare, Finland, firstname.lastname@thl.fi

Nonresponse is a ever-progressing issue with population surveys. However, population-based surveys are a vital source of information, because registers contain only limited range of information. By combining survey and register data and using effective missing data methods it is possible to gain more reliable results using survey data suffering from nonresponse issues.

In 2010 Finnish Regional Health and Well-being Study (ATH survey) was conducted by National Institute for Health and Welfare with a national sample of 5,000 and regional samples from Kainuu (9,000), Northern Ostrobothnia (8,000) and city of Turku (9,000). The response rate varied from 37 % to 65 % between different age groups and regions. Overall in the whole data set (n= 31,000) the response rate was 49 %.The study included questions mainly concerning health and well-being. We used this data to compare different missing data handling methods such as Inverse probability Weighting (IPW, used for unit nonresponse), weighted sequential Hot Deck imputation (WSHDI) and multiple imputation (MI) which work with MAR or MNAR data. When using these methods it was also possible to see how they affect the survey results compared to results using only sample weights (CC analysis).

As an outcome in this work we used the self-reported depression which had one of the highest item nonresponse rates in the survey. The question was "Have you had any of the following conditions diagnosed or treated by a doctor over the past 12 months?" and depression was one of the alternatives in the question. Because of the question structure and disposition of the alternative, also valuable information was gained from the response indicators of the previous and the following alternative. The overall item nonresponse rate for the self-reported depression was 9.4 % among the participants of the survey. The outcome prevalences of self-reported depression were studied in two age groups 25 to 64-year-olds and over 65 year-olds.

To gain information of the nonrespondents and to develop better nonresponse models we used also register data from National Register Center (sample variables such as age, gender and marital status), Statistics Finland (education and profession of the respondent) and Social Insurance Institute (special reimbursement of medication such as psychotropic medication). The register data was available for the full sample.

Overall the nonresponse bias corrected results of self-reported depression prevalences were quite similar in CC analysis, IPW and WSHDI. MI rates stand out significantly especially in over 65-year-olds, which shows that also the non-complete predictors affect the results. All the methods seemed to respect the relation between areas and age groups compared to CC analysis but also compared to the register-based rates of people receiving reimbursement of depression medication. Overall the results indicate that CC analysis underestimates the rate of depression in every area studied. Correcting this nonresponse bias is vital especially if there is further analysis to be done with depression in the analysis model. The results also suggest that the willingness to answer the depression question depends on depression itself which means that the data is MNAR and the nonresponse corrected estimates may still be biased. Also further examination especially of the MI predictor sets should be done.

References

Kaikkonen R., Murto J., Pentala O., Koskela T., Virtala E., Härkänen T., Koskenniemi T., Ahonen J., Vartiainen E. & Koskinen S. Results of Regional Health and Well-being Study 2010-2014. Internet publication: www.thl.fi/ath. 2010-2014.

Härkänen T., Kaikkonen R., Virtala E. and Koskinen S. Inverse probability weighting and doubly robust methods in correcting the effects of non-response in the reimbursed medication and self-reported turnout estimates in the ATH survey. doi:10.1186/1471-2458-14-1150. BMC Public Health 2014, 14:1150.

Pentala O. Väestötutkimusaineiston tilastolliset kadonhallintamenetelmät - Alueellisen terveys- ja hyvinvointitutkimuksen kyselyaineisto 2010 (Statistical Nonresponse Methods in Population Surveys - Regional Health and Wellbeing study 2010). Master's Thesis. University of Helsinki. http://hdl.handle.net/10138/144287, 2014.

BAYESIAN SUBCOHORT SELECTION FOR LONGITUDINAL COVARIATE MEASUREMENTS IN FOLLOW-UP STUDIES

Jaakko Reinikainen
Department of Mathematics and Statistics, University of Jyväskylä, Finland,
jaakko.o.reinikainen@jyu.fi

Juha Karvanen
Department of Mathematics and Statistics, University of Jyväskylä, Finland

A fundamental question in study design is how to answer the research question as precisely as possible with a limited budget. We consider planning longitudinal data collection in follow-up studies where covariates are time-varying. We assume that the entire cohort cannot be selected for longitudinal measurements due to financial limitations and study how a subset of the cohort should be selected optimally in order to obtain precise estimates of covariate effects in a survival model. In our approach, the study will be designed sequentially utilizing the data collected in previous longitudinal measurements as prior information. We propose using a Bayesian optimality criterion in the subcohort selections, which is compared to simple random sampling with simulated and real follow-up data. This study extends our previous results, where optimal subcohort selection was studied with only one re-measurement and one covariate, to more realistic cases where several covariates and measurement points are allowed. Our results show that the optimal subcohort includes individuals with a high risk of an event and, on the other hand, with extreme covariate values. The results support the conclusion that the precision of the estimates can be clearly improved by optimal design.

SYNTHETIC DATA SOURCES IN THE SPATIAL ANALYSIS OF POVERTY IN POLAND

Wojciech Roszka
Department of Statistics, Poznan University of Economics, POLAND, wojciech.roszka@ue.poznan.pl

One of the key elements of the state's social policy is the poverty prevention. Information about its spatial diversity is very helpful in the context of allocation of resources and to take measures to prevent its growth. Characteristics linked to the poverty estimates are mainly based on European Union Statistics on Income and Living Conditions (EU - SILC), which is a sample survey conducted annually in all countries of the European Union. Limitations associated with the sample size mean that the at-risk-of-poverty rate are estimated at most to the level of the provinces (NUTS 2, voivodeships in Poland).
The application of the small area estimation methods (SAE) allows better quality estimation without increasing the sample. The problem in the context of poverty mapping in Poland was, among others, raised by Wawrowski (2014), who used a Fay - Herriot model to estimate the at-risk-of-poverty rate at NUTS 3 level. Simultaneously with the model approach methods, techniques for creating synthetic full-coverage datasets on the basis of available information are being developed (Haslett et al. 2010). These techniques are already of interest to the European Union in the context of the spatial diversity of the living conditions of the population (Alfonso et al. 2011).
The synthetic datasets are being created using dynamically developing methods of spatial microsimulation. They allow to create multi-dimensional estimates for small domains. Especially synthetic reconstruction methods will be presented. With the joint usage of statistical matching methods (Raessler 2002) and iterative proportional fitting models (Rahman 2008) it is possible to construct the synthetic micro-populations at a small area level in such a way that all known constraints at the small area level are reproduced.
The aim of the paper is to assess the possibility of obtaining multivariate estimates of acceptable quality for at-risk-of-poverty at NUTS 4 (poviats in Poland) using a synthetic data set created on the basis of EU - SILC 2011 dataset and information from the census. The resulting estimates will be evaluated by their reliability, consistency and quality, as well as a comparative analysis will be carried out with the results obtained by "classical" SAE models.

References

Alfonso A., Filzmoser P., Hulliger B., Kolb J-P., Kraft S., Munnich R., Templ M., 2011, *Synthetic Data Generation of SILC Data*, European Commision, Community Research, AMELI Project
Haslett S., Jones G., Noble A., Ballas D., 2010, *More or Less? Comparing small area estimation, spatial microsimulation, and mass imputation*, Section on Survey Research Methods JSM, American Statistical Association
Raessler S., 2002, *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Springer, New York, USA
Rahman A., 2008, *A Review of Small Area Estimation Problems and Methodological Developments*, Discussion paper 66, NATSEM, University of Canberra
Wawrowski ., 2014, *Wykorzystanie metod statystyki maych obszarw do tworzenia map ubstwa w Polsce* [*The use of small area estimation methods for poverty mapping in Poland*], Wiadomoci Statystyczne 9/2014, Polskie Towarzystwo Statystyczne [*Polish Statistical Association*]

MEAN ESTIMATION WITH ROBUST CALIBRATED ESTIMATORS

Iryna Rozora
Taras Shevchenko National University of Kyiv, Ukraine, irozora@bigmir.net

Olga Lukovych
Taras Shevchenko National University of Kyiv, Ukraine, lukolga@ukr.net

Calibration weighting is a general technique for adjusting probability-sampling weights to increase the precision of estimates to be consistent. Let's, however, focuses on the basic conditions: single phase sampling and full response. In practice, survey conditions are not that simple, but many theory papers nevertheless address this situation.
In Rozora et al. (2014) mean estimation with calibration approach was studied. In this paper we continue the investigation of different calibration techniques to estimate mean.
A trimmed mean, as a statistical measure of central tendency, is also considered. The main advantage of the trimmed mean is robustness and higher efficiency for mixed distributions and heavy-tailed distribution (like the Cauchy distribution), at the cost of lower efficiency for some other less heavily-tailed distributions (such as the normal distribution). The algorithm for construction of trimmed mean estimators using calibration techniques (calibrated trimmed mean) is given.
As an alternative way to estimate the population mean in the case of symmetric the median of Walsh averages is considered. The median of Walsh averages has a robustness property and is more efficient, for example, in the case of normal distribution, than trimmed mean. We propose to consider new estimator for the population mean that is based on the using of auxiliary information, calibration techniques and the median of Walsh averages.
The properties of these estimators are studied and natural population example is also considered.

References

Rozora I., Lukovych O. and Stovba V. (2014) Mean Estimation Calibration Approach in Survey Sampling. Workshop of BNU Network in Survey Statistics, Tallinn, Estonia, 95-102.

M.Rueda, S.Martinez, A.Arcos, and J.F.Munos (2009) Mean Estimation Under Successive Samplinng with Calibration Estimators. Communication in Statistics – Theory and Method, 38, 808-827.

C.-E. Särndal (2007). The calibration approach in survey theory and practice. Survey methodology, 33(2), 99-119.

AN APPLICATION OF UNIT-LEVEL MODEL FOR FRACTIONS OF UNEMPLOYED

Tomas Rudys
Vilnius University Institute of Mathematics and Informatics, Lithuania, tomas.rudys@mii.vu.lt

We are interested in estimating the fractions of unemployed for small areas using Lithuania Labor Force survey data. In practice often happens that the estimates for small areas are needed after the realization of the sample, that is during the construction of sample design not enough attention was given to get adequate sample size in all small areas. Later, using the design-based estimators and having not sufficient sample sizes in small areas, this leads to get estimates of an not adequate precision. In this work we are using model-based approach to get small area estimates for fractions of unemployed. Very important role in this case plays the auxiliary information which could be available at unit level or at more aggregated levels. The use of unit level auxiliary information may lead to get more precise estimates for small areas. Here we are using two-phase unit-level model ([1,2,3]) to get the estimates for fractions of unemployed. Model checking is done by analyzing how model fits the data. An inference to unknown model parameters is done by using Bayesian approach. We are comparing Bayesian and frequentist approaches for estimating unknown model parameters by modeling. For the Bayesian inference the R package LaplacesDemon is used ([4]).

References

1. H. J. Boonstra, B. Buelens, and M. Smeets. *Estimation of municipal unemployment fractions – a simulation study comparing different small area estimators,* Statistics Netherlands, Projectnr: DMH-205714 (2009).

2. P. J. Farrell. *Bayesian inference for small area proportions*, Sankhya: The Journal of Statistics, 2000, Volume 62, Series B, Pt. 3, pp. 402-416.

3. J. N. K. Rao, *Small area estimation,* Hoboken: John Wiley & Sons (2003).

4. Statisticat LLC (2014). *LaplacesDemon: Complete environment for Bayesian Inference*. R package version 14.04.05, URL http://www.bayesian-inference.com/software.

IMPROVING OF THE RELIABILITY OF UKRAINIAN POVERTY INDICATORS ESTIMATES USING AUXILIARY INFORMATION

Volodymyr Sarioglo
Ptoukha Institute for Demography and Social Studies of the National Academy of Science of Ukraine, Ukraine, sarioglo@idss.org.ua

The state Household Living Conditions Survey (HLCS) provided by the State Statistics Service of Ukraine on quarterly basis is the main source of information for measuring of a number of important indicators which in details reflect incomes, expenditures, consumption features, poverty of Ukrainian households and many others. In Ukraine, there are 24 oblasts, one autonomous republic (Autonomous Republic of Crimea – AR of Crimea) and two cities (the city of Kyiv and the city Sevastopol) which are all consisted as administrative regions (at present AR of Crimea, city Sevastopol and partly two eastern regions are temporally occupied by Russia). Each administrative region is subdivided into lower level administrative units. Analysis of reliability of the poverty indicators estimates defined on base of HLCS data, first of all of the relative poverty rate by national poverty line which is defined as 75% of median household equivalence income, has proved the indicators direct estimates at the regional level to be insufficiently reliable. The most important factors of this are the high nonresponse rate and small effective sample size in some regions which can make less than 250 households. Respectively, this limits the possibilities of efficient poverty monitoring in Ukraine, especially in the context of government decentralization policy implementation, and affects the income and consumption distributions which are determined from the survey. The high nonresponse rate registered among the territories with relatively high rate of well-off population (primarily the capital Kyiv and other cities) make the possibility to suggest about incomplete coverage of well-to-do households by the HLCS.

An efficient approach to enhancing the reliability of the poverty indicators estimates at the regional level is using the small area estimation technique (Longford, 2005; Longford, 2010). The possibility of the adequate procedures development and their further practical usage depends essentially on the available additional information. According to results of our researches today one of the most effective approaches in Ukraine in addition to indirect estimation is calibration (Deville, J.-C. and Särndal, C.-E., 1992) of HLCS statistical weights using the structure of the NAS final consumption expenditures by groups of goods and services in order to ensure maximum proximity of the HLCS structure to NAS structure. It should be noted that it is necessary to eliminate the main differences between expenditures measures in NAS and in survey taking into account the methological features.

On base of the obtained results it is shown that the reliability of poverty indicator estimates in Ukraine can be enhanced by using of NAS data on household final consumption expenditures on different stages of the indicators estimation process. The calibration of statistical weights using NAS data can decrease biases of estimates for regions as it provides the possibility of better accounting of well-to-do households' expenditures by the HLCS data. The indirect estimation using NAS data and poverty indicators direct estimates for the national level and composite ones for the regional level from previous year can significantly decrease the mean squared error of estimation.

References

Longford N. T. (2005) Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician. Springer-Verlag, New York.
Longford N.T. (2010) Simulation of small-area estimators of the poverty rates in the oblasts of Ukraine. - SNTL and UPF, Barcelona, Spain. The report prepared for the Social Assistance System Modernization Project, Ukraine, Kyiv.
Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87:376–382.

GENERALIZED SOLUTIONS FOR DATA EDITING AT SURS

Rudi Seljak
Statistical Office of the Republic of Slovenia, rudi.seljak@gov.si

Kaja Malešič
Statistical Office of the Republic of Slovenia, kaja.malesic@gov.si

Data editing is a process aimed at providing accurate, complete and consistent information. In general, it encompasses all activities related to the detection of errors, correction of inconsistencies and imputation of missing values in the observed data. Due to its complexity, it is mostly very time and resource consuming. It is therefore rational for statistical organizations to harmonize different editing processes within a variety of surveys with appropriate general technical solutions.

The paper presents the main characteristics of the new approach to data editing implemented at the Statistical Office of the Republic of Slovenia (hereinafter SURS) with the use of a generalized software tool and its impact on the statistical process. This tool, the so-called MetaSOP application (SOP is an acronym in Slovenian for statistical data processing), will encompass also other parts of data processing (e.g. aggregation and standard error estimation, tabulation and tabular protection, quality indicators). At the moment the module for data editing is already in production, while other modules are being tested or developed. The editing module combines three IT environments: ORACLE process metadata database, SAS macros as general programs for data processing and .Net WPF application. The whole system is based on a metadata driven principle. Its core is a general program code consisting of SAS macros for logical checks, deterministic, systematic and individual corrections and imputations. For usage in a particular survey, this general code is parameterized with appropriate metadata entries.

After the implementation of the new editing approach into the process of first surveys we have already noted the first advantages, such as elimination of errors in consistency between entered rules and variables. We expect other positive results, such as rationalization of statistical processes and overall improvement in data quality (also due to different organizational procedures and expected increased number of surveys having the automated editing), to be visible later. The goal is that all surveys at SURS will be gradually included in the generalized process.

Keywords: *metadata driven systems; statistical data editing; generalized software solutions.*

References

Seljak, R. (2009). Integrated statistical systems and their flexibility – How to find the balance?. NTTS Conference, Brussels, Belgium, 5-7 March, 2013.

Seljak R. (2014). Metadata driven application for data processing – from local toward global solution. Conference of European Statisticians. Paris, France, 28-30 April, 2014.

A DIFFERENT IMPUTATION APPROACH FOR THE CATEGORICAL SURVEY STUDIES

Alper SİNAN
Sinop University, Turkey, alpsin@sinop.edu.tr

Aydın KARAKOCA
Necmettin Erbakan University, Turkey, akarakoca@konya.edu.tr

Missing data is a common problem for all areas of statistics. Especially, in the survey analysis, researchers are often faced with the problem of nonresponse. Thus, the statistical inferences of surveys are not reliable in such cases. Various imputation methods and different analysis techniques are developed for dealing with the missing data. These methods have different approaches which depend on the type of the data.
In this paper, a new approach is developed for dealing with missing data in a categorical survey study. Suggested method base on finding a subset of similar respondents to the respondent with missing data. Imputation is made randomly according to the probabilities which determined from the chosen subset. The application of suggested method is made with a real survey data and the results of this categorical imputation approach are interpreted.

References

Beale, E.M.L., and Little R.J.A., 1975, Missing Values in Multivariate Analysis, Journal of the Royal Statistical Society. Series B (Methodological), 37, 1, 129-145.

Craig K. Enders, 2010, Applied Missing Data Analysis, The Guilford Press, New York, United States of America.

Finch, W.H., 2010, Imputation Methods for Missing Categorical Questionnaire Data: A Comparison of Approaches Journal of Data Science 8, 361-378

Little, R. J. A., and Rubin, D. B., 2002, Statistical analysis with missing data, John Wiley.

THE DEVELOPMENT OF PRODUCTION COSTS IN DAIRY FARMS USING PANEL DATA

Alina Sinisalo
Natural Resources Institute Finland (LUKE), Finland, alina.sinisalo@luke.fi

The development of production costs was studied by using the annual accounting data from dairy farms (specialized in dairying) taking part in Luke profitability bookkeeping for the years 2000—2013. The dataset was formed as panel allowing the possibility to effectively study the change over time. Annual accounting data from each farm was used. Data set was unbalanced since it is voluntary to participate in Luke bookkeeping activities and, on the other hand, some farms had exited the business. During the period 2000—2013 the number of dairy farms has dropped by 57 per cent, but the number of cows by 22 per cent. The production costs of farms grew overt the studied period not only due to increased input costs but also due to the fact that the average farm size had grown over time. The unit production costs was studied by using unbalanced panel data with a linear mixed model taking into account farm-level information and time effect. Production costs increase year-to-year. The unit cost decreased as the number of cows increased. Small farms had higher unit cost and also annual variation was larger than medium-sized and large farms.

References

Sinisalo, A. 2015. Production costs of Finnish dairy farms in the 2000s. Proceedings of the 2015 International Conference “ECONOMIC SCIENCE FOR RURAL DEVELOPMENT” No37, Jelgava, LLU ESAF, 23-24 April 2015, pp. 26-34. Available at: http://www.esaf.llu.lv/journals-and-proceedings.

TESTING DIFFERENCES OF MEANS

Alina Sinisalo
Natural Resources Institute Finland (LUKE), Finland, alina.sinisalo@luke.fi

Arto Latukka
Natural Resources Institute Finland (LUKE), Finland, arto.latukka@luke.fi

Anne-Mari Sepponen
Natural Resources Institute Finland (LUKE), Finland, anne-mari.sepponen@luke.fi

In Finland the profitability bookkeeping of farms has long traditions of more than 100 years. Accounting and other farm-level data are collected every year from more than 1000 voluntary farms. By suitably weighting the results with weight factors calculated individually for each farm taking into account the type of operations, economic size and location by support areas, the data is used to describe the results of all Finnish farms. The results are published on Economy Doctor online service.

On Economy Doctor online service various descriptive statistics of agricultural enterprises are reported on average-level by different classification factors. These averages may be different between groups, but user is unable to distinguish are the differences significant and which groups differ from each other. A system of analyses shall be produced that automatically analyses differences between groups and provides reporting accordingly. The aim is to develop Economy Doctor online service even more versatile and to produce concrete help for researchers, decision makers and advisers.

We compare and test the means of different variable groups - are the differences statistically significant. If statistically significant differences are found, the differences are real not random and test results can be seen in the reports of EconomyDoctor as typical statistic report ($p<0,05$ *, $p<0,01$ **, $p<0,001$ ***). If the tests prove that means are different from each other statistically significant we also test differences of means pairwise.

The system is under construction. Necessary codes and logical chains including various parametric and non-parametric tests have been written in SAS. Written code first checks the number of observations and required assumptions before statistical testing. How to visually and technically report the results on Economy Doctor is under consideration.

References

EconomyDoctor Internet-service: http://www.mtt.fi/economydoctor

Analysis of RSU International Brand – Views by International Students

Inga Skendere
Riga Stradins University, Latvia, Inga.skendere-dregere@rsu.lv

Julija Stare
Riga Stradins University, Latvia, jūlija.stare@gmail.com

In Europe the dynamics of gradually increasing student international mobility even more accentuates the study internationalisation aspects not only at the level of education policy, but also study process provision level. These tendencies consequently lead to reinforced evaluation of the existing practice and point to potential development direction for a systematic and targeted improvement of the study quality. One of the most crucial preconditions for development provision is a regular student needs analysis.

The aim of this research is evaluate the experience of Rīga Stradiņš University in pedagogical work with international students in order to identify concurrent accomplishments and opportunities for growth. In the framework of the research student views about the study quality at Rīga Stradiņš University have been outlined with the help of quantitative and qualitative research methods. The research results allow to draw a conclusion that RSU is presented as a recognizable higher education brand at an international level.

Keywords: brand, higher education space, internationalization, intercultural dimension.

CHALLENGES IN INTEGRATING ADMINISTRATIVE VAT DATA INTO UK SHORT-TERM OUTPUT STATISTICS

Markus Gintas Šova
Office for National Statistics, UK, markus.sova@ons.gov.uk

The integration of administrative VAT data into short-term output statistics offers the prospect of reducing both respondent burden and survey costs. However, such administrative data raises some new challenges and compounds some existing challenges to survey data. In the UK about 90% of VAT-registered enterprises report their VAT turnover on a quarterly basis according to one of three reporting schedules (each with a different set of starting months). Apart from the issue of calendarisation, there is a consequent impact on the time it takes to receive complete (or near-complete) data. The effect of this response delay is exacerbated by a striking relationship between data timeliness and reported turnover. There are also differences in the definitions of turnover used for statistical and tax purposes. In some instances there are issues of data quality, without the opportunity to verify surprising data values with the reporting enterprise. Data can be revised, sometimes several times and sometimes by more than an order of magnitude.
This paper examines some recent research undertaken in the UK to address these methodological issues.

RESIDENCE TESTING USING REGISTERS – CONCEPTUAL AND METHODOLOGICAL PROBLEMS

Ene-Margit Tiit
Statistics Estonia and University of Tartu

Different register are nowadays used more and more often in official statistics and research. Also register-based censuses that have since been a luxury for a few highly advanced countries will be organized in more countries.

That means some new conceptual problems, for instance – what is the population size of a acountry?

In general, since the population size has been calculated by census. If the census is over- or undercovered, must the population size of census be corrected?

What is the population size in the case of register-based population? If the population register is over- or undercovered, should the number be corrected?

In Estonia a set of administrative register has been used for estimating the undercoverage of census 2011. Also it has been checked if the same methodology works for estimating the census population for register-based census in 2020.

The following task is to elaborate the registers-based algorithm for regular estimating the current population size for population statistics.

WHAT CAN BE LEARNED ABOUT SURVEY NON-RESPONSE THROUGH RECORD LINKAGE? EXAMPLES FROM HEALTH EXAMINATION SURVEYS.

Hanna Tolonen
National Institute for Health and Welfare, Finland, hanna.tolonen@thl.fi

Juha Karvanen
University of Jyväskylä, Finland, juha.karvanen@jyu.fi

Päivikki Koponen
National Institute for Health and Welfare, Finland, paivikki.koponen@thl.fi

Erkki Vartiainen
National Institute for Health and Welfare, Finland, erkki.vartiainen@thl.fi

Kari Kuulasmaa
National Institute for Health and Welfare, Finland, kari.kuulasmaa@thl.fi

Declining participation rates are a problem in survey research, especially since data is not missing at random. Survey non-participants differ from participants in characteristics related to the outcomes of interest. In Europe, the participation rates in health examination surveys (HESs) used to be around 80-90% in 1970's and 1980's. In many recent HESs participation rates have been 40-60%.[1,2]

In Finland, we have a unique possibility to link survey samples to several nationally representative administrative registers using personal identification codes. In the framework of the Non-participation in Health Examination Survey (NoPaHES) project (http://www.ehes.info/nopahes), we have linked several Finnish health examination surveys (FINRISK study 1972-2012, Health 2000/2011 surveys, and Migrant Health and Wellbeing Survey) to national administrative registers covering causes of death, hospital discharges, cancers, entitlement to specifically reimbursed medications due to specific conditions, and purchase of medications by Anatomical Therapeutic Chemical **(**ATC) codes, socio-economic position and geographical information of the places of residence and the examination centres. This allows us to study the characteristics of survey non-participant in respect to their socio-economic position and health outcomes, and to learn about the effect of non-participation on the quality and representativeness of the survey results. We have observed that survey non-participants have excess mortality in 1-20 years follow up. Non-participants also have higher hospitalization rates during the survey period than participants.

Data obtained via record linkage offers also a unique opportunity to compare methods for missing data analysis with real data and to develop new methods that utilize the data available for the non-participants.

References

[1] European Health Interview & Health Examination Surveys Database. Available at https://hishes.wiv-isp.be
[2] Tolonen H, Ahonen S, Jentoft S et al. Differences in participation rates and lessons learned about recruitment of participants – The European Health Examination Survey Pilot Project. Scand J Public Health 2015; 43(2):212-9.

OUTLIER DETECTION METHODS FOR BUISENESS SURVEYS

Anton Tovchenko
State Statistics Service of Ukraine, Ukraine, A.Tovchenko@ukrstat.gov.ua

Olexiy Tkachenko
State Statistics Service of Ukraine, Ukraine, tom@ukrstat.gov.ua

The quality of sample enterprise survey, unlike sample social survey, greatly depends upon the number of outliers and the methods of outlier detection (that is, enterprises with big values of relevant indicator). Enterprise statistics has a number of traits:

1. The distribution of observed value is not normal, it is hyperbolic, and so the values of about 5% of enterprises can make up to 90% of total value. In addition, it makes using parametric methods unfeasible (e.g. Pirson correlation or dispersion analysis).
2. In enterprise statistics the stratified sampling is usually used (stratum=domain). The accuracy of estimation of domains is of critical importance. In addition, the sum of values across different domains may greatly vary (10 times or more).
3. While building sample design, the real values are usually known across all the general population, because the census usually precedes the sample survey.

Due to mentioned traits, there is a question as to how many outliers to detect and with which method. In addition, the not only the accuracy of estimation of total population must be assured, but the estimation of all the relevant domains as well.

The conclusion on the quality of the sample was based on the value of dispersion of estimation of total and coefficient of variation of estimation, total and across domains. The quantity of enterprises in the sample of about 20% of the total population was deemed necessary. The testing of the design consisted of the following steps:

1. Outlier detection individually for each domain.
2. Allocation of the rest of the enterprises (total quantity minus the quantity of outliers) across domains using Neumann allocation.
3. Sampling.
4. Calculation of coefficient of variation of estimation (total and for each domain).

State Statistics Service of Ukraine specialists tested the following methods of outlier detection:

1. Iterative parametric method: "n-sigma" coefficient for standard deviation was changed from 1 to 5 by step 0,2.
   (parameter > mean + coefficient*(standard deviation))
2. Iterative non-parametric method: coefficient for interquartile range was change from 1 to 5 by step 0,2.
   (parameter > median + coefficient*(interquartile range))
3. Iterative minimization of coefficient of variation in each domain to a value that was changed form 300% to 100% by 10%.
   (sort by relevant indicator in descending order, iteration: "calculate CV, if CV is greater than specified value - top enterprise is an outlier", continue until CV becomes lesser or equal to specified value)

Preliminary results:

1. Using parametric method leads to the worst quality (especially in some domains), what is more, the optimal coefficient for standard deviation varies in range from 1 to 2, which is contrary to "heuristic" three.

2. Using non-parametric methods gives better results across domains; the optimal coefficient for interquartile range is in range from 4 to 5.
3. Using Iterative minimization of coefficient of variation in each domain gives the best results.

The presentation and the paper will illustrate and describe the results with the conclusion of which method was better to detect outlier.

References

Kish L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: John Wiley & Sons.

Hansen M.H., Hurvitz W.N. & Madow W.G. (1953). *Sample survey methods and theory*. New York: John Wiley & Sons.

CHILD CARE CHOICES IN FINLAND: COPING WITH INCOMPLETE REGISTER-BASED DATA

Maria Valaste
The Social Insurance Institution, Finland, maria.valaste@kela.fi

The Finnish Government plans to cut the subjective right to municipal child day care from parents who are at maternity leave, paternity leave, parental leave or child home care allowance. After parental leave (at which point child is 9–10 months of age) until the child is 3 years of age there is available: subjective right to a municipal child day care or cash for care, which is called child home care allowance.

Analyses of potential outcomes of a possible reform may assessed with a microsimulation model (Haataja and Valaste 2014). Our analysis makes use of the Finnish static microsimulation model SISU. Microsimulation is an empirically based data modelling technique that has been traditionally used in the areas of taxation, social benefits and other types of economic activity.

Register-based data are originally produced for other purposes than for specific research question (Wallgren and Wallgren 2007). Our challenge for analysis is that there is information about children's care spells in public day care but the information is inadequate. Very little systematic work has been done to validate the data or to produce systematic imputation or editing for various abnormal observations (Haataja and Juutilainen 2012). This is an attempt to utilize child based spell data. We aim to utilize existing data sources to construct a new in-home and out-of-home child day care model which would incorporate information from the perspectives of children, their parents and the family as a whole.

The data covers 50 percent of mothers who gave birth in 1999-2009 in Finland, their spouses and children. The data comprises Kela benefits and background information. Our aim is to place care spells into a monthly calendar for each child under school age using existing data sources. To combine information on different spells, a priority system has been created (Haataja et al. 2013). The priority system produces a single status for a child for each month in the calendar year. The idea is that the most reliable or logical data source is selected first, in the absence of that the next, and so on. Imputation rules have been implemented to solve the problem of missing data in the child day care spells. We present the outcomes of simulations relating to the current reforms and comparisons of the attendance of public day care between our calendar data and the other data sources.

References

Haataja, A. and Juutilainen, V-P. (2012). Päivähoitotietoa Kelassa. Nettityöpapereita 36. Helsinki: Kelan tutkimusosasto.

Haataja, A., Mattila, J., and Valaste, M. (2013). Applying individual level data on children's care periods to microsimulation models. http://hdl.handle.net/10138/154062

Haataja, A., and Valaste, M. (2014). Applying child-based information to a microsimulation model. A better tool to assess outcomes of alternative entitlements to child care provisions?. Working papers 52. Helsinki: Kela Research Department.

Wallgren, A., and Wallgren, B. (2007). Register-based statistics: administrative data for statistical purposes. John Wiley & Sons.

A SURVEY OF STUDENT SURVEYS

Olga Vasylyk
Taras Shevchenko National University of Kyiv, Ukraine, ovasylyk@univ.kiev.ua

This academic year the first group of Master students specialized in statistics, which had been studying in accordance with new programme and therefore had possibility to learn an advanced course of "Sample surveys theory and methods", graduated the Faculty of Mechanics and Mathematics of Taras Shevchenko National University of Kyiv. Three students from this group decided to write master theses in survey statistics. I suggested them to make real surveys of students studying at the Faculty of Mechanics and Mathematics. They were free to choose any topic, which seemed interesting for them. Planning a survey, preparation of a questionnaire, selection a sample, interviewing and processing of data, and, finally, analysis of sample data were supposed to be made by a student alone. In order to minimize the influence of a scientific supervisor, there were no limitations or instructions concerning sampling design as well as estimation methods and software for data analysis. My objective was to understand (basing on a student's preferences and difficulties), what issues were complicated for students, and to identify deficiencies in the program of the course in survey statistics or gaps in the lecture materials.

As a result, the following surveys were conducted: "Living in dormitory" by Julia Brateshko, "Why I entered the Faculty of Mechanics and Mathematics" by Victoria Martynenko, and "Topicality of the diploma awarded by the Faculty of Mechanics and Mathematics" by Iryna Kramar. In each survey several characteristics were investigated, and the results are very interesting. For instance, students are not satisfied with living conditions in their dormitories, but at the same time they are not willing to help staff of the dormitories to make these conditions more comfortable. As for decision to enter the faculty of mathematics, the majority of students were influenced by their parents. And the survey of graduates showed that mostly employment does not correspond to their speciality, but studying at this faculty helped them much. More details about surveys' results will be presented at the conference.

Target populations in these surveys were not identical, because in the first one only students living in dormitories were surveyed, in the second survey the population consisted of all students studying at our faculty now, and in the third survey graduates from the faculty in 2012, 2013 and 2014 were surveyed. Nevertheless, in all three surveys students used stratified simple random sampling with stratification on the year of study or graduation. For estimation of parameters of interest the Horvitz-Thompson estimator was used. Nonresponses were ignored, although methods of weighting and imputation were presented at lectures. As characteristics of accuracy, coefficients of variation and design effects were calculated. Two students analyzed their data in SPSS and one used R. When it comes to relationships among variables, the students love cross tabulation and do not like regression.

Many questions arose during preparation of these theses including those, which seemed obvious for me, but appeared to be not so obvious for students. But my main conclusion is that it would be better to increase number of lectures and practical lessons on such topics as analysis of data with missing values, regression analysis and regression estimators, as well as usage of additional information in general. And more attention should be given to detailed explanations WHY in some situations it is important or necessary to use special techniques.

References

1. Brateshko, Ju. *Sample survey "Living in dormitory"*. Master theses, 2015.
2. Kramar, I. *Sample survey "Topicality of the diploma awarded by the Faculty of Mechanics and Mathematics"*. Master theses, 2015.
3. Martynenko, V. *Sample survey of students of the Faculty of Mechanics and Mathematics "Why I entered the MechMat"*. Master theses, 2015.

METADATA OF THE EUROPEAN TIME USE SURVEY DATABASE

Paavo Väisänen
Statistics Finland, Paavo.Vaisanen@stat.fi

Meta data of the database of the Harmonized European Time Use Surveys contain the description on the applied statistical methods, estimation processes, nonresponse adjustment, imputation, data collection, coding process, reliability, timeliness, confidentiality and the harmonisation of activity codes. European Time Use Surveys were harmonised by Eurostat in 2000 and the new updated guidelines of the Harmonised European Time Use Surveys (HETUS) were published in 2009 (Eurostat, 2009). Twenty countries have announced to be interested to deliver the data of their time use survey into the HETUS database. The need of a more detailed standard structure for the collection and dissemination of quality reports has grown up during last years since there was no homogeneity between the existing structures used in the different statistical domains. As a part of the European Statistical System (ESS) was developed structured tool, the Metadata Handler (ESS-MH), to help the production, management, exchange and dissemination of metadata within Eurostat and international organizations. ESS-MH is a web based application for reference metadata production, exchange and dissemination in the ESS (see Eurostat Info Space). The ESS Standard Quality Report Structure is the main report structure for reference metadata related to data quality. The statistical agencies of the participating countries are asked to fill in the metadata work flows of the ESS-MH attached with detailed methodological and quality descriptions as appendixes.
Eurostat granted financial resources to comprise the database of micro data and Statistics Finland collects the micro data from countries and constructs the database. The responsibility of Statistics Finland is to harmonize the data files and collect the meta data of the data base. The meta data contains classifications, variable definitions and descriptions from the application of statistical methods that are used in local time use surveys. Meta information concerning the applied statistical methods is appended to the Data Base to help researches to understand possibilities and restrictions caused by the structure of the sample design, the estimation procedures and the data collection. Eurostat calculates statistical tables from the database and publishes them. There will not be public access to the HETUS database but the metadata are published in internet.

References

Eurostat (2009). Harmonised European Time Use Surveys: 2008 guidelines. Luxembourg: Office for Official Publications of the European Communities

Eurostat Info Space, EU Comission. SDMX and Metadata Standards, https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/European_Statistical_System_Metadata_Handler_%28ESS_MH%29

AN EIGENPROBLEM APPROACH TO OPTIMAL EQUAL-PRECISION SAMPLE ALLOCATION IN SUBPOPULATIONS

Jacek Wesołowski
CSO, Warsaw, Poland

In a survey often the constraints for precision of estimators of subpopulations parameters have to be taken care of during the allocation of the sample. Such issues are often solved with mathematical programming procedures. It is desirable to allocate the sample, in a way which forces the precision of estimates at the subpopulation level to be both: optimal and comparable, while the constraints of the total (expected) size of the sample (or samples, in two-stage sampling) are imposed. We show that such problem in a wide class of sampling plans has an elegant mathematical and convenient computational solution involving eigenvalues and eigenvectors of matrices defined in terms of some population quantities. As a final result we present a simple method for calculating the subpopulation optimal and equal-precision allocation which is based on one of the most standard algorithms of linear algebra (available e.g. in R software).

Theoretical solutions are illustrated through a numerical example based on the Labour Force Survey. The method allows to accommodate rather automatically for different levels of precision priority for subpopulations.

This is a joint work with R. Wieczorkowski (CSO, Warsaw, Poland).

5 PARTICIPANTS

Family name	First name	Affiliation	Email
Alho	Juha	University of Helsinki	juha.alho@helsinki.fi
Andersson	Per Gösta	Stockholm University	per.gosta.andersson@stat.su.se
Antoni	Manfred	Institute for Employment Research (IAB)	manfred.antoni@iab.de
Beresewicz	Maciej	Poznan University of Economics	maciej.beresewicz@ue.poznan.pl
Berger	Yves	University of Southampton	Y.G.Berger@soton.ac.uk
Bethlehem	Jelke	Leiden University	jelkeb@xs4all.nl
Bobrova	Anastacia	Institute of Economics of NAS of Belarus	nastassiabobrova@mail.ru
Bokun	Natallia	Belarus State Economic University	natalliabokun@rambler.ru
Bondarenko	Iana	Dnipropetrovsk National University	iana.s.bondarenko@gmail.com
Budkina	Natalja	Riga Technical University	natalja.budkina@rtu.lv
Chebanova	Mariia	Taras Shevchenko National University of Kyiv	M_Chebanova@ukr.net
Dirdaite	Ieva	Vilnius Gediminas Technical University	dirdaite.ieva@gmail.com
Djerf	Kari	Statistics Finland	kari.djerf@stat.fi
Folasade	Ariyibi	Office for National Statistics, UK	fola.ariyibi@ons.gsi.gov.uk
Galiautdinovaite	Kristina	Vilnius Gediminas Technical University	kristina.galiautdinovaite@gmail.com
Hedlin	Dan	Stockholm University	dan.hedlin@stat.su.se
Härkänen	Tommi	THL	tommi.harkanen@thl.fi
Ianevych	Tetiana	Taras Shevchenko National University of Kyiv	yakovenkot@gmail.com
Julin	Päivi	Aalto university	julin@apup.org
Karakoca	Aydin	Necmettin Erbakan University	akarakoca@konya.edu.tr
Karvanen	Juha	University of Jyväskylä	juha.t.karvanen@jyu.fi
Katajamäki	Esa	Natural Resources Institute Finland	esa.katajamaki@luke.fi
Keto	Mauno	University of Jyväskylä	mauno.keto@mamk.fi
Kilinc Kan	Betul	Anadolu University	bkan@anadolu.edu.tr
Kojalo	Una	Riga Stradins University	Una.Kojalo@rsu.lv
Kokkinen	Mirva	Natural Resources Institute Finland	mirva.kokkinen@luke.fi
Kopra	Juho	University of Jyvaskyla	juho.j.kopra@jyu.fi
Krapavickaite	Danute	Vilnius Gediminas Technical University	danute.krapavickaite@vgtu.lt
Kuoppa-aho	Mika	Natural Resources Institute Finland	mika.kuoppa-aho@luke.fi
Laaksonen	Seppo	University of Helsinki	seppo.laaksonen@helsinki.fi
Lahiri	Parthasarathi	JPSM, University of Maryland	plahiri@umd.edu
Laiho-Kauranne	Johanna	Natural Resources Institute Finland	johanna.laiho-kauranne@luke.fi
Larchenko	Anna	Ministry of Foreign Affairs, Belarus	annalarchenko@gmail.com
Lehtinen	Jaana	University of Helsinki	jaana.lehtinen@helsinki.fi
Lehtonen	Risto	University of Helsinki	risto.lehtonen@helsinki.fi
Lepik	Natalja	University of Tartu	natalja.lepik@ut.ee
Li	Yan	University of Maryland	YLI6@UMD.EDU
Liberts	Martins	Central Statistical Bureau of Latvia	martins.liberts@gmail.com
Louhivaara	Anna	Natural Resources Institute Finland	anna.louhivaara@luke.fi
Lumiste	Kaur	University of Tartu	kaur.lumiste@ut.ee
Lysa	Olha	Ptoukha Institute for Demography and Social Studies of the National Academy of Science of Ukraine	Olysa@ukr.net
Lähteenmäki	Mervi	Helsingin yliopisto	mervi.lahteenmaki@helsinki.fi
Maasing	Ethel	Statistics Estonia	ethel.maasing@stat.ee
Malesic	Kaja	Statistical Office of the Republic of Slovenia	kaja.malesic@gov.si
Malmila	Mia	City of Vantaa	mia.malmila@vantaa.fi
Masiulaityte-Sukevic	Inga	Statistics Lithuania	inga.masiulaityte@stat.gov.lt
Moilanen	Pentti	Natural Resources Institute Finland	pentti.moilanen@luke.fi
Moretti	Angelo	University of Manchester	angelo.moretti@postgrad.manchester.ac.uk
Möttönen	Jyrki	University of Helsinki	jyrki.mottonen@helsinki.fi

Neuvonen	Marjo	Natural Resources Institute Finland (Luke)	marjo.neuvonen@luke.fi
Nurminen	Markku	University of Helsinki, MarkStat Consultancy	markku.nurminen@markstat.net
Oinonen	Saara	Statistics Finland	saara.oinonen@stat.fi
Oksanen	Pihla	University of Helsinki	pihla.oksanen@helsinki.fi
Ollila	Pauli	Statistics Finland	pauli.ollila@stat.fi
Pahkinen	Erkki	University of Jyväskylä	pahkinen@maths.jyu.fi
Pentala	Oona	National Institute for Health and Welfare	oona.pentala@thl.fi
Piela	Pasi	Tilastokeskus	pasi.piela@tilastokeskus.fi
Piliutsik	Andrei	Institute of Economics of NAS of Belarus	pilutic@gmail.com
Pohjanpää	Kirsti	Statistics Finland	kirsti.pohjanpaa@stat.fi
Reinikainen	Jaakko	University of Jyväskylä	jaakko.o.reinikainen@jyu.fi
Rendtel	Ulrich	Freie Universität Berlin	ulrich.rendtel@fu-berlin.de
Roszka	Wojciech	Poznan University of Economics	wojciech.roszka@ue.poznan.pl
Rota	Bernardo	Örebro University	dsitao@yahoo.com
Rozora	Nataliia	Nielsen	rozora@ukr.net
Rozora	Iryna	Taras Shevchenko National University of Kyiv	irozora@bigmir.net
Rudys	Tomas	Vilnius University	tomas.rudys@mii.vu.lt
Ruotsalainen	Kaija	Statistics Finland	kaija.ruotsalainen@stat.fi
Sarioglo	Volodymyr	Ptoukha Institute for Demography and Social Studies of the National Academy of Science of Ukraine	sarioglo@idss.org.ua
Schulte Nordholt	Eric	Statistics Netherlands	e.schultenordholt@cbs.nl
Sinan	Alper	Sinop University	alpsin@sinop.edu.tr
Sinisalo	Alina	Natural Resources Institute Finland (Luke)	alina.sinisalo@luke.fi
Skendere	Inga	Riga Stradins University	inga.skendere-dregere@rsu.lv
Slickute-Sestokiene	Milda	Statistics Lithuania	milda.slickute@stat.gov.lt
Sõstra	Kaja	Statistics Estonia	kaja.sostra@stat.ee
Sova	Markus Gintas	Office for National Statistics (UK)	markus.sova@ons.gov.uk
Taskinen	Pertti	Statistics Finland	pertti.taskinen@stat.fi
Tiit	Ene-Margit	Statistics Estonia, University of Tartu	ene.tiit@ut.ee
Tolonen	Hanna	National Institute for Health and Welfare	hanna.tolonen@thl.fi
Tovchenko	Anton	State Statistical Service of Ukraine	A.Tovchenko@ukrstat.gov.ua
Traat	Imbi	University of Tartu	imbi.traat@ut.ee
Tuhkuri	Joonas	ETLA, The Research Institute of the Finnish Economy	joonas.tuhkuri@etla.fi
Valaste	Maria	The Social Insurance Institution of Finland, KELA	maria.valaste@helsinki.fi
Vasylyk	Olga	Taras Shevchenko National University of Kyiv	olva75@gmail.com
Vehkalahti	Kimmo	University of Helsinki	kimmo.vehkalahti@helsinki.fi
Veijanen	Ari	Statistics Finland	ari.veijanen@stat.fi
Väisänen	Paavo	Statistics Finland	paavo.vaisanen@stat.fi
Wesolowski	Jacek	Central Statistical Office of Poland	intrelations@stat.gov.pl
Würbach	Ariane	Leibniz Institute for Educational Trajectories	ariane.wuerbach@lifbi.de
Zhang	Li-Chun	University of Southampton & Statistics Norway	L.Zhang@soton.ac.uk
Zimoch	Urszula	University of Helsinki	urszula.zimoch@helsinki.fi