LATVIJAS UNIVERSITĀTE
UNIVERSITY OF LATVIA
ANNO 1919

LATVIJAS STATISTIKA

LSA
LATVIJAS STATISTIĶU ASOCIĀCIJA

Vidzemes Augstskola
UNIVERSITY OF APPLIED SCIENCES

WORKSHOP OF BALTIC-NORDIC-UKRAINIAN NETWORK ON SURVEY STATISTICS

August 24-28, 2012
Valmiera, Latvia

University of Latvia
Central Statistical Bureau of Latvia

WORKSHOP OF BALTIC-NORDIC-UKRAINIAN NETWORK ON SURVEY STATISTICS

August 24-28, 2012
Valmiera, Latvia

Organised by University of Latvia, Central Statistical Bureau of Republic of Latvia,
Association of Latvian Statisticians, Vidzeme University of Applied Science

LECTURE MATERIALS AND CONTRIBUTED PAPERS

Riga, 2012

Workshop of Baltic-Nordic-Ukrainian Network on Survey Statistics. August 24-28, 2012, Valmiera, Latvia. University of Latvia, Central Statistical Bureau of Latvia, Riga, 2012, 232 p.

Baltijas-Ziemeļvalstu-Ukrainas seminārs par apsekojumu statistiku. 2012.g. 24.-28.augusts, Valmiera, Latvija. Latvijas Universitāte, Latvijas Republikas Centrālā statistikas pārvalde, Rīga, 2012, 232 lpp.

Programme Committee

Danutė Krapavickaitė, (Vilnius Gediminas Technical University, Statistics Lithuania)
Gunnar Kulldorff (University of Umeå)
Jānis Lapiņš (Bank of Latvia)
Risto Lehtonen (University of Helsinki, chair)
Mārtiņš Liberts (Central Statistical Bureau of Latvia, University of Latvia)
Aleksandras Plikusas (Institute of Mathematics and Informatics, Vilnius University)
Daniel Thorburn (University of Stockholm)
Imbi Traat (University of Tartu)
Olga Vasylyk (National Taras Shevchenko University of Kyiv)

Organizing committee

Andris Fisenko (Central Statistical Bureau of Latvia)
Jānis Lapiņš (Bank of Latvia)
Mārtiņš Liberts (Central Statistical Bureau of Latvia, University of Latvia, chair)
Inta Priedola (TNS Latvia)
Biruta Sloka (University of Latvia)

Organizers

Central Statistical Bureau of Latvia
University of Latvia
Association of Latvian Statisticians
Vidzeme University of Applied Science

Sponsors

The Nordplus Programme of the Nordic Council of Ministers
International Association of Survey Statisticians (IASS)
Central Statistical Bureau of Latvia

Layout and compilation: Harijs Kārkliņš

ISBN 978-9984-45-557-0

Preface

The Baltic-Nordic co-operation on survey statistics started in 1992 initiated by Professor Gunnar Kulldorff. A Baltic-Nordic network for co-operation on education and research in survey statistics was established in 1996. The network was expanded in 2008. It is called the Baltic-Nordic-Ukrainian Network on Survey Statistics now.

The network is organising annual events as summer schools, workshops or conferences since 1997. The Workshop of Baltic-Nordic-Ukrainian Network on Survey Statistics organised in August 2012 in Valmiera, Latvia is the 16th event in the series. We are expecting 63 participants at the workshop. It is the highest number of participants at the workshops organised by the network. The participants represent twelve countries.

Carl-Erik Särndal (Sweden) and Monica Pratesi (Università di Pisa, Italy) are the main speakers of the workshop. They will give three lectures each. Danutė Krapavickaitė (Vilnius Gediminas Technical University, Statistics Lithuania), Gunnar Kulldorff (Umeå University, Sweden), Pauli Ollila (Statistics Finland), Ulrich Rendtel (Freie Universität Berlin, Germany), Anders Wallgren (Sweden) and Li-Chun Zhang (Statistics Norway) are the invited speakers of the workshop. Each invited speaker will give one lecture to the workshop participants.

Most of the other workshop participants will present contributed papers. 42 contributed papers are accepted for presentations. A poster session is arranged for the first time at the workshops organised by the network. More information about the workshop is available at the workshop website (home.lu.lv/~pm90015/workshop2012).

I express thanks to all members of the Organising Committee and Programme Committee for the active involvement in the organisation of the workshop. Special thanks are due to Uldis Ainārs, Ance Ceriņa, Harijs Kārkliņš, Ilga Puisāne, Renāte Rudzīte, Jeļena Vaļkovska, Kaspars Vasaraudzis, Līga Zalužinska, Aija Žīgure and other staff members of the Central Statistical Bureau of Latvia who are helping in the organisation of the workshop. Finally I would like to thank Nordic Council of Ministers (The Nordplus Programme), International Association of Survey Statisticians, the Central Statistical Bureau of Latvia and University of Latvia for a given support.

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

Riga, July 2012

Mārtiņš Liberts

Contents

Main Speakers

Main speakers and lecture titles:

Prof. Carl-Erik Särndal

I. Interplay between survey theory and the demands of official statistics production, a theory of science perspective

II. The data collection stage: Responsive design and balancing the set of respondents

III. The estimation stage: Calibrated weighting for nonresponse bias reduction and preferably without increased variance

Prof. Monica Pratesi

I. Recent developments in small area estimation (SAE) methodology

II. Use of SAE in Italy: Case studies

III. SAMPLE Project – data integration and SAE software

Invited Speakers

Invited speakers and lecture titles:

Dr. Doc. Danutė Krapavickaitė

Sampling methods used in the studies of natural resources

Prof. Gunnar Kulldorff

Twenty Years of Baltic-Nordic Co-operation – With Expansion to Ukraine and Belarus

Dr. Pauli Ollila

Process model for statistical editing

Prof. Dr. Ulrich Rendtel

Teaching Survey Statistics by Teleteaching: A joint project at three German universities (abstract)

Anders Wallgren

Administrative registers, survey system design and quality assessment

Dr. Scient. Li-Chun Zhang

Micro calibration for data integration

Invited papers

Process Model for Editing

Pauli Ollila[1]

[1]Statistics Finland, e-mail: pauli.ollila@stat.fi

Abstract

This paper outlines a process model for editing with more detailed description of its phases and premises in statistical production.

Keywords: Editing, imputation, process model

1 Introduction

Statistical data editing is essential for achieving sufficient data quality needed for the production of statistics. Missing values and all kinds of errors in data, incoherencies between variables and in time, exceptional distributions, various sources of information and challenging calculations are some aspects which should be considered during editing the data. The process of error detection, correction and imputation has been very heterogeneous, often time and resource consuming and in many cases not so systematic and consistent over time.

An essential part of the modernization of statistical production is expressing the production in terms of a process, e.g. the work for the *generic statistical business process model* within UNECE (Vale, 2011) and in some statistical offices. One attempt to formulate editing in a process form has been made by Luzi et al. (2007). Recently there has been international activity as well (Zhang, 2011).

The common editing process is applicable for several statistics of different kind, providing a framework for more standardized and efficient actions of editing. The process includes main phases and possibilities to iterative actions depending on the information gained during different phases of data collection and treatment. The realization of the process requires *methodological choices* suitable for the situation as well as *decisions for proceeding in the process*. The process can be controlled and guided with the *estimates* and *quality indicators* obtained during different phases of the process. In order to be useful, the process should be *supported by a suitable IT environment* with proper software for realizing the methodological decisions and required practices and collecting metadata for further actions and quality control. The process should be *in harmony with other standardized management systems* and the *general production process of the statistics*.

Statistical data editing reacts to a vast variety of problems occurring in statistical data. In the course of time the practices of editing have been very heterogeneous and occasionally non-systematic, ineffective and inconsistent. This situation might have caused quality, resource, cost and timeliness problems. The standardization of statistical production by constructing a common editing process provides solutions for these kinds of problems. Moreover, many countries develop editing in a data environment utilizing tax registers and other source data provided by the administration. This multi-source data situation requires a process of editing which also takes the special nature of administrative data into account.

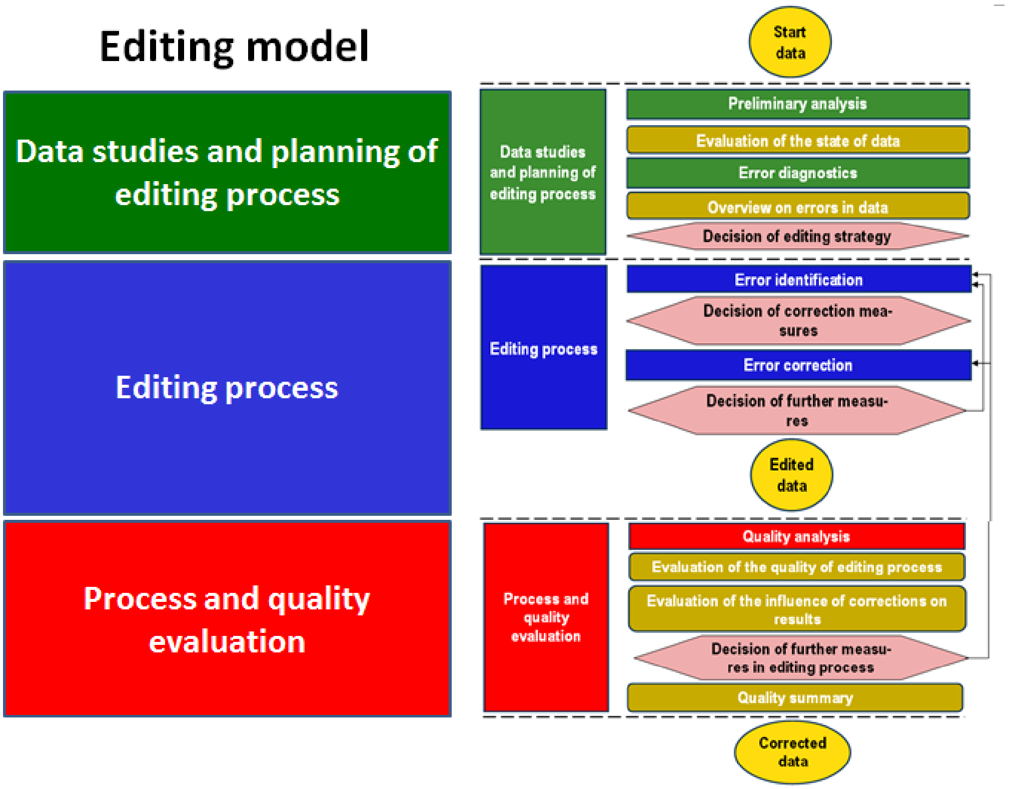
This paper is based on the work of the editing project at Statistics Finland, closing at the end of 2011 (Ollila & Rouhuvirta, 2011). The model is still in the draft phase and it is subject to changes. The English terms used in this paper are not final and they will be reviewed later.

2 Process Model for Editing

2.1 Main Structure

The **process model for editing** includes three main phases: *data studies and planning of editing process*, *editing process*, *process and quality evaluation*. The term "editing" is used here in a broad sense, and it includes actions connected to both recognition and correction of errors. Every main phase is consists of *action entities*, *evaluations* and *decisions*. Figure 1 presents the model at the general level. The **action entity** consists of *actions* (not seen in Figure 1) targeted to the data to be used for statistics. These **actions** cause changes in the data and provide new information for use (new variables and descriptive information). The **evaluation** is made by the researcher or another person connected to the statistics, and it can be aimed to the data in the process, the results from the actions and/or actions in the process. The **decision** of the researcher defines forthcoming actions. The action entities include evaluations and decisions as well. The essential feature of the model is the possibility to go back in the process. The phases are dealt with in more detail in the subsequent sections.

Figure 1: Process Model for Editing


Editing model
Data studies and planning of editing process
Editing process
Process and quality evaluation
Start data
Data studies and planning of editing process
Preliminary analysis
Evaluation of the state of data
Error diagnostics
Overview on errors in data
Decision of editing strategy
Editing process
Error identification
Decision of correction mea-sures
Error correction
Decision of further measu-res
Edited data
Process and quality evaluation
Quality analysis
Evaluation of the quality of editing process
Evaluation of the influence of corrections on results
Decision of further measu-res in editing process
Quality summary
Corrected data

2.2 Data Studies and Planning of Editing

2.2.1 Preliminary analysis

The **preliminary analysis** gives *an overview on the substance state of current data*, which might be raw data or partially processed data. The preliminary analysis includes two subphases: *data analysis based on prepared programs* and *interactive data study*.

The **data analysis based on prepared programs** includes tabulation and calculation of statistics with relevant subgroups targeted to variables essential for editing process. The basis for this phase can be well-chosen tabulation practices from the previous rounds. Some estimates can be defined as ”State of data” indicators, which can be calculated at the subsequent phases as well for evaluating the development of editing (resembling Canada’s ”rolling estimates”, Saint-Pierre & Bricault, 2011). The contents of the programs should be quite constant from one round to another providing tabulations and results, which would enable *comparison between rounds*. On the other hand, when new error phenomena occur, the programs should be updated. The variables and the error situations depend on the data, but for these aims there should be generic programs (e.g. macros, modules), which allow the required constant form easily.

Interactive data study is interactive analysis based on the experiences of the researcher using suitable IT solutions (analysis methods, graphical methods, observation value views). The aim is to catch those (possibly new) characteristics, which cannot be found with prepared programs or when further studies are needed based on suspicious results from the prepared program studies.

2.2.2 Error Diagnostics

In the **error diagnostics** phase the goal is to make an *overview on typical errors in the data* and possible *changes in the error profile of the data*. As a separation from the *error identification* phase in the actual editing process, here the *error identification and further actions due to that are not the goal*, though in some cases the errors could be identified. The error diagnostics includes the *error analysis based on prepared programs* and *the interactive error study*.

The **error analysis based on prepared programs** includes tabulations of fatal errors and clear suspicions found in the data. The variables in the programs, their classifications and the estimators to be used must be decided before realizing program runs. As in data analysis, the contents of the programs should be quite constant for comparison. On the other hand, when new error phenomena occur, the programs should be updated. The variables and the error situations depend on the data, but for these aims there should be generic programs (e.g. macros, modules), which allow the required constant form easily.

The **interactive error study** is (as in preliminary analysis phase) interactive analysis based on the experiences of the researcher using suitable IT solutions (analysis methods, graphical methods, observation value views). At this phase the goal is to find errors (e.g. systematic), which could not be revealed with previous error procedures. The proceeding of the study and the choices of various study tools depend on the results. This study should be continued until a sufficient level is reached.

2.2.3 Deciding the Editing Strategy

Based on the preliminary analysis one can make an **evaluation of the state of the data**. The evaluation can include estimates (including specified “state-of-data” indicators) and other tabulations from prepared programs and statistics, graphical products, listings and tables from the interactive data study. Correspondingly, the product of the error diagnosis is an evaluation of the error situation in the data including the same kind of information as mentioned above. It does not include exact observational and variable-level error identifications.

These evaluations together with the definitions of the starting point of the process model (see Chapter 3) made by the persons conducting the statistics and judgments of previous experiences and practices form the basis for the **decision of the editing strategy**. It includes a preliminary outline: what actions are realized, in what order and with what criteria (parameters), when also taking into account the constraints of the data. The plan can be specified or changed due to information gained during the editing process. For some statistics with less complicated and rarely changing structure the preliminary analysis and the error diagnosis probably consist of only few operations.

2.3 Editing Process

2.3.1 Overall Level

All editing (broad definition) is realized in the phase of **editing process**. It consists of the *error identification*, the *decision of correction measures*, the *error correction* and the *decision of further measures*. It can include various actions of error identification and error correction and it is *iterative*, i.e. either by following the strategy of the editing process or by changing the plan to some extent due to new information the researcher chooses the methods, how to proceed in the editing process. The string of actions of error identification and error corrections can be called an *"editing path"*. These paths can vary from very simple operations to complex systems with a lot of constraints.

2.3.2 Error Identification

The **error identification** phase includes actions, which result as a whole to identifying errors in certainty (i.e. fatal errors) and possible errors at the observation level or at the group of observations level, including non-structural missing values. The decisions come from the previous phase, i.e. *data studies and planning of editing strategy*. Part of the actions might describe possibility of error in general or in suspicious subsets (e.g. macro editing) or tell that something is wrong in the observation, but the error is not identified. These require further actions, but they are a vital part of a process which ends to a situation where there are one or more observations and their variables sufficiently identified for corrections. The term *"error detection"* is used more often than *"error identification"* in the literature and articles (see e.g. De Waal *et al.*, 2011), but here this choice emphasizes that before moving to error corrections observations and variables have full error identifiability.

The error identification phase provides information in different forms. This information is presented in various **views**, which might be printed into a paper form in some cases. An *information view* is any kind of form of information presented on the screen of the computer. Most of the final decisions for error identifications are based on the researcher evaluation of this provided information. The classification of views here is: the **single unit view** (part or all of the variable values of one observation can be seen), the **data view** (the matrix with observations and their variable values can be seen), the **observation list** (listing with limitations in observations and/or variables), the **calculation table** (the presentation due to a tabulation in the data), the **result or statistics list** and the **graphics** in various forms.

In order to get the views which are needed, one has to **conduct realizations** of these views with suitable software tools. The choices of views for different editing situations, the planning of the visualization, the choices of procedures, modules etc., and the ways of realizing these views during the process can affect the efficiency and quality of editing and decision-making.

In the **automatic identification** there is no evaluation based on views, but the identified error information moves straight to the error correction phase, where the corrections are made according to exact predefined rules.

The **non-processed identification** brings to the view only the statistical data and possibly some reference or auxiliary variables from other sources (e.g. previous values). In practice this kind of identification happens only with the one observation view or the data view, and then the decision of the error is based on the "overall look on the data" or comparison by the researcher.

The **processed identification** includes processing of the data and possibly some reference or auxiliary data to new variables or analysis on various levels. The outcomes are new variables in the observations, calculated statistics and/or analytical quantities, which should help the error identification. The processing for error identification is divided here into three categories: *edit rules, analytic processing, macro level processing*.

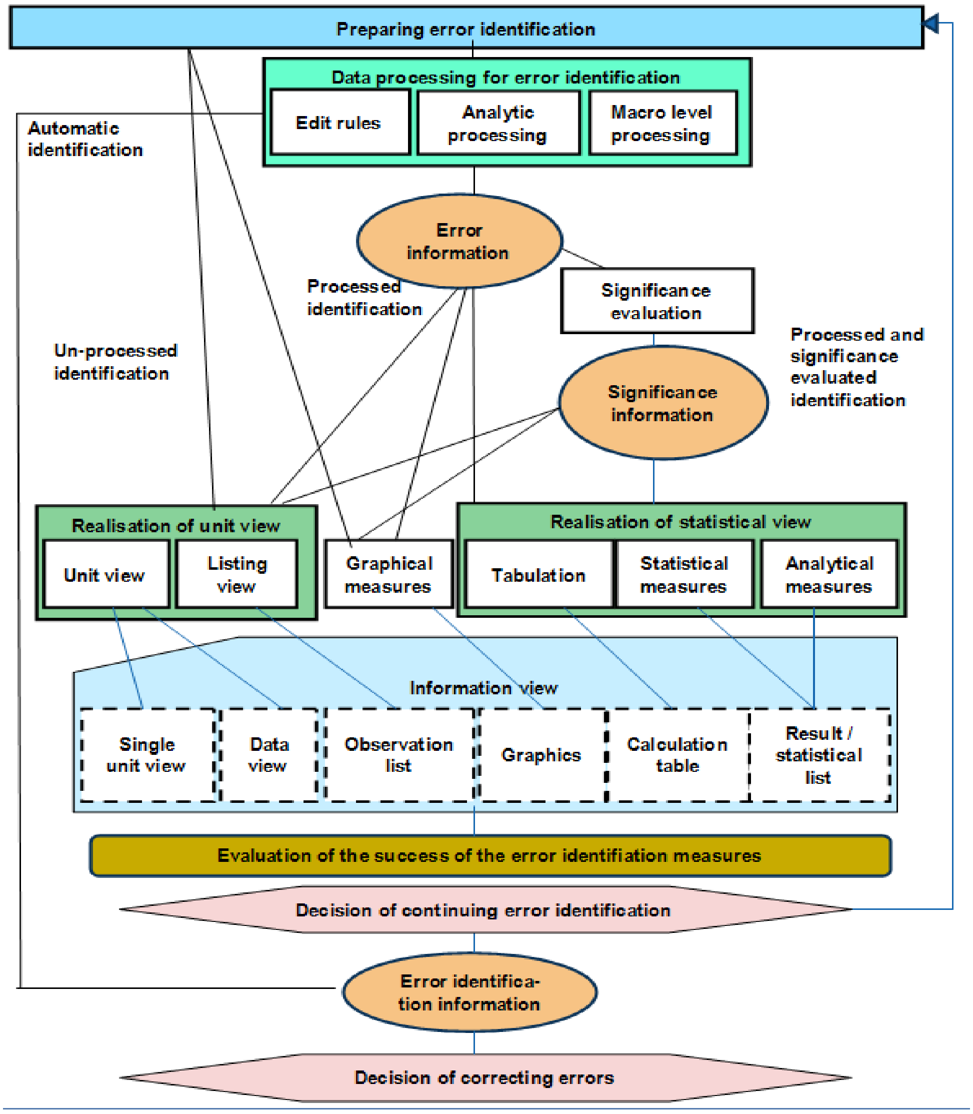
The **edit rules** are logical conditions connected to variables, their functions or external information. With the edit rules one can recognize errors at the observation level. Some edit rules recognize errors with certainty (fatal errors), but it is rather usual that suspicious values are found with some limit values for variables or simple functions of them. The main idea is just to indicate that the rule is or is not fulfilled. In simple cases the observations are listed based on edit rules, and quite often a separate indicator variable for that edit rule is created for further use. Some edit rules can be *constraints*, which are required in the data for some variables.

The **analytic processing** applies all kinds of statistical and mathematical methods to the observations in order to reveal errors. The most common outcomes of these operations are analytical quantities, e.g. distance measures for assessing outliers (see De Waal *et al.*, 2011) or the method by Hidiroglou and Berthelot (1986) based on ratio quantities in time), predicted values based on modelling for editing purposes (see De Waal *et al.*, 2011) or methods for error localization in edit rules using reliability weights (e.g. De Waal *et al.*, 2011).

The starting point of the **macro level processing** is the calculation of statistics at the data or subset level (often coinciding to real results and subsets as well). The aggregate level study is a rather common practice in statistics making. These results are compared in time, connected to the reference results available at the moment or some functions are made from the results (e.g. ratio). The macro level processing provides information about possible state of error at the data or subset level, and thus the real identification of errors must be conducted in subsequent operations, guided by the results from the macro level.

The **processed and significance evaluated identification** includes the study of significance of the variable values and observations to the results, usually expressed with scores (see Hedlin, 2008). It is possible to deal with more variables at once and evaluate the total score for observation as well. These actions direct time consuming interactive studies (manual editing) to a limited set of most influential observations, leaving the rest of the observations to quick correction routines or uncorrected. This practice called selective editing is considered to improve efficiency and save resources and expenses (Adolfsson & Gidlund, 2008).

Figure 2: Phase of Error Identification


Preparing error identification
Data processing for error identification
Edit rules
Analytic processing
Macro level processing
Automatic identification
Error information
Processed identification
Significance evaluation
Un-processed identification
Processed and significance evaluated identification
Significance information
Realisation of unit view
Unit view
Listing view
Graphical measures
Realisation of statistical view
Tabulation
Statistical measures
Analytical measures
Information view
Single unit view
Data view
Observation list
Graphics
Calculation table
Result / statistical list
Evaluation of the success of the error identifiation measures
Decision of continuing error identification
Error identifica-tion information
Decision of correcting errors

2.3.3 Error Correction

The **error correction** phase realizes corrections of all or some identified errors following the decisions made at the error identification phase. This broad definition includes imputation as well. It is possible, that some identified error is not corrected, because it is decided to be negligible or the correction method is not justified (e.g. too few observations, not enough evidence for a choice of any methodological corrections). Figure 3 shows the structure of the error correction phase. The methods of error correction are divided into two classes: *non-methodological* and *methodological corrections*. In these classes there can be noticed a division between *searched* and *created values*.

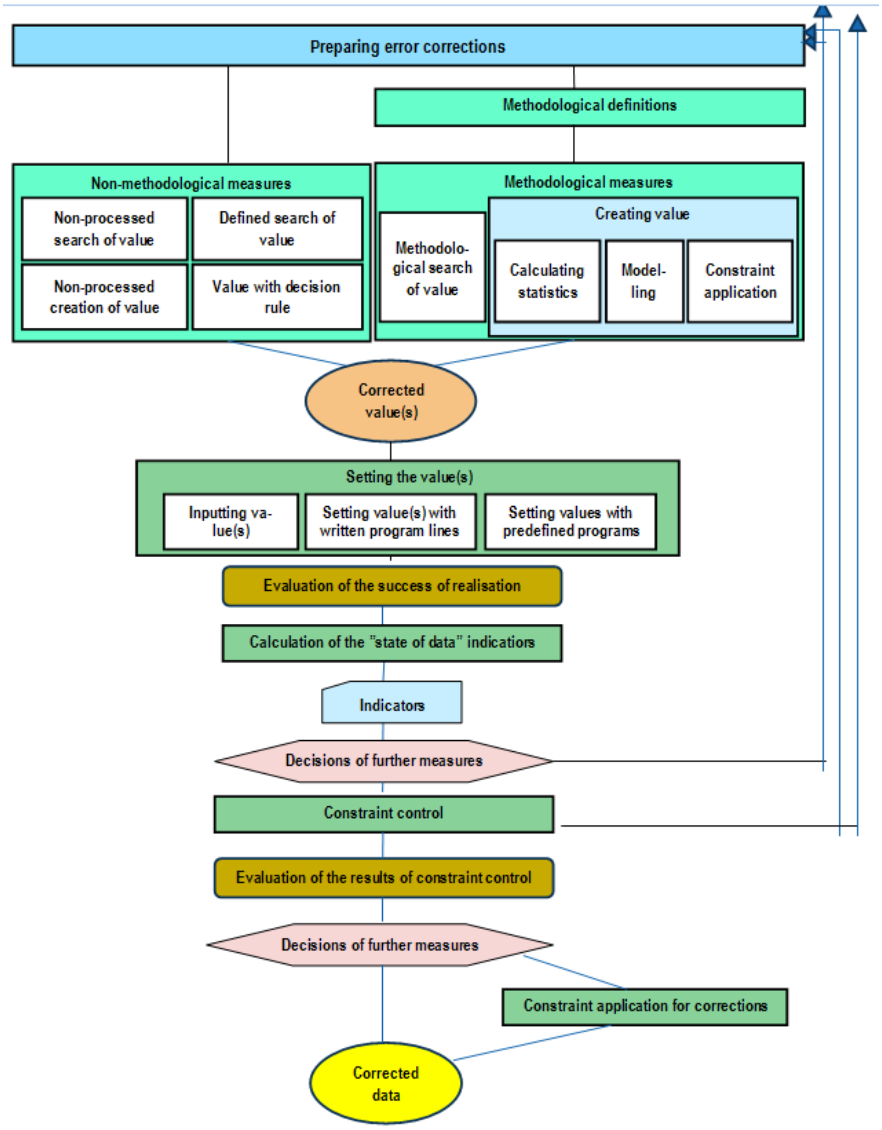
The **non-methodological corrections** are: *non-processed search of value, non-processed creation of value, defined search of value* and *value with decision rule.* The **non-processed search of value** is usually a value obtained with a contact to the value provider or the respondent; sometimes the contact reveals that the erroneous value is right. The value can also be fetched manually from another source (e.g. publication, register). The **non-processed creation of value** is a researcher's decision of the value which should be used, based on some reasoning. The **defined search of value** includes a programmed value search mechanism, targeted to some data of a previous round (rather usual) or a data including auxiliary information (e.g. register, another source data or statistics). Obtaining the **value with a decision rule** is a common way to correct an error with discrete variables.

The **methodological corrections** are: *methodological search of a value, calculating statistics, modelling* and *constraint application.* Often correcting with these practices is called imputation, though there are broader definitions as well. The methodological search of value is conducted through donor set of observations. This set is usually restricted to some subset of respondents. The observations with item non-response or errors obtain values from a donor chosen with some methodological principle. It can vary from random selection to some function-controlled donor methods (e.g. distance measures for a variable which exist for all or nearly all observations). In **calculating statistics**, the calculations targeted to the whole data or to a subset provide a statistics (e.g. mean or median). Correspondingly, in modelling one creates a regression model or another kind of model for predicted values used for imputation, sometimes with a stochastic residual added. The **constraint application** corrects the error or missingness fulfilling the requirements of a constraint (e.g. sum of subtotals = given overall total). It is either some function of existing values in the constraint or with full item response in a constraint some smoothing function of all or some values.

The methodological correction methods may require **definitions** e.g. for parameters, limit values or information needed for the successful conduct of the method. Also non-methodological methods can include some definitions.

The corrected values should be put into the data. Three alternatives of **setting the values** are provided here: **inputting value(s)** via a unit view (for many statistics a constructed application), **setting value(s) with written program lines** (surprisingly common alternative, might be unavoidable in very complex situations with several corrections) and **setting values with predefined programs** (might be e.g. software modules [*Banff* imputation procedures or *Selekt* macros] or programs possibly controlled with process parameters). The success of this realization is evaluated. The "state of data" indicators are calculated after the corrections. If the realization has not been successful, one must go back to specify or alter the correction process or in rare cases leave the correction phase (in some difficult item non-response cases). After that the next error identification can be conducted or if there are constraints connected to the variables in correction, one can proceed to the **constraint control**. It ensures that the constraints are satisfied in the data. If not, then last constraint applications are conducted. After this one can get back to error identification or one can consider the editing process ended. The result of this is the **corrected data**.

Figure 3: Phase of Error Correction


Preparing error corrections
Methodological definitions
Non-methodological measures
Non-processed search of value
Defined search of value
Non-processed creation of value
Value with decision rule
Methodological measures
Methodolo-gical search of value
Creating value
Calculating statistics
Model-ling
Constraint application
Corrected value(s)
Setting the value(s)
Inputting va-lue(s)
Setting value(s) with written program lines
Setting values with predefined programs
Evaluation of the success of realisation
Calculation of the "state of data" indicators
Indicators
Decisions of further measures
Constraint control
Evaluation of the results of constraint control
Decisions of further measures
Constraint application for corrections
Corrected data

2.4 Process and Quality Evaluation

Process and quality can be evaluated with indicators, which should be calculated automatically at least when the data is considered to be at the final stage, but also when the data and the processing is in such a situation that an evaluation of what has been done is needed. The process of calculation should be in a constant form. . There are several indicators defined form process and quality evaluation (e.g. Euredit, 2004, Luzi *et al.*, 2007, Eurostat, 2009, and Ollila, 2012).

The *"state-of-data" indicators* (essential estimates at the population level and in relevant subgroups, as in preliminary analysis and during editing process) were discussed earlier, but they are applicable here as well. The progress of these indicators may bring valuable information about the changes during the process.

The **indicator describing the editing process** is a statistic, which enables the study of actions for error identifications or error corrections. Some examples are the *edit failure rate* (Eurostat, 2009), i.e. "the proportion of responding units for which an error signal is triggered by a specified checking algorithm", the *number of observations failing at least one edit rule* (Luzi *et al.*, 2007) or the *imputation rate* (Luzi *et al.*, 2007, Eurostat 2009).

The **indicator revealing the influence of editing on results** is a statistic, which enables the study of the change of estimates due to the editing process. Some examples are the *weighted relative average imputation impact* (Luzi *et al.*, 2007) and the *weighted imputation error ratio* (Luzi *et al.*, 2007).

The **indicator in relation with previous results** is a statistic, which reveals the effect of the editing process in estimates when compared with the previous round. A simple example is the *relative change of estimates between two time points.*

3 Premises of the Process Model for Editing

The process model is based on three main contributors: *personnel of the statistics, methodologists and IT experts*. The statistics should define information required by the editing model (variables and criteria for them, constraints, variables for indicators, requirements for process). The methodologists should provide resources for the model, i.e. the methodology bank, the concept library and instructions for actions and decisions at different phases. The IT experts should define and plan solutions of process information required by the editing model (saving information of E&I actions and indicator calculation). Further, suitable software (e.g. Banff, Selekt, LogiPlus, SAS JMP) are integrated for the phases of the modules. When needed, new modules could be created.

The actions realized in the editing model are supported with the knowledge included in the **methodology bank**, which describes the methods included in the methodology groups in the different phases of the editing model. **Method** as a term can be considered here broadly: in addition to *statistical, mathematical* and *logical actions* it includes *consistent courses of actions*. The structure of the methodology bank follows strictly the **methodology groups** appearing in the editing model.

Table 1: Methodology groups in the model

Measures describing data	Refining data	Search of value	Setting value	Creating value
Realization of unit view	Edit rules	Non-processed search of value	Inputting value	Non-processing creation of value
Realization of listing view	Analytic processing	Defined search of value	Setting values with written program line	Value with decision rule
Calculation of statistical measures	Macro level processing	Methodological search of value	Values with predefined programs	Value with calculating statistics
Realization of tabulation	Significance evaluation			Value with modelling
Realization of analytical measures				Value with constraint application
Realization of graphics				

References

Adolfsson, C. & Gidlund, P. (2008). *Conducted Case Studies at Statistics Sweden*. Supporting paper in UNECE meeting, Vienna, Austria.

De Waal, T., Pannekoek, J. & Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Wiley.

Eurostat (2009). *ESS Standard for Quality Reports*, Eurostat, Luxemburg.

Hedlin, D. (2008). *Local and Global Scores in Selective Editing*. Invited paper in UNECE meeting, Vienna, Austria.

Hidiroglou, M. & Berthelot. J. (1986). Statistical editing and imputation for periodic business surveys. *Survey methodology*, **12**.

Luzi, O., Di Zio, M., Guarnera, U., Manzari, A., De Waal, T., Pannekoek, J., Hoogland, J., Tempelman, C., Hulliger, B.&, Kilchmann, D. (2007): *Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys*, EDIMBUS project report.

Ollila, P. & Rouhuvirta, H. (2011). *Process Model for Editing (draft)*, Internal methodology paper (in Finnish), Statistics Finland.

Ollila, P. (2012). *Indicators of Raw Data, Editing and Imputation (draft)*, internal methodology paper (in Finnish), Statistics Finland.

Saint-Pierre, E. and Bricault, M. (2011). *The Common Editing Strategy and the Data Processing of Business Statistics Surveys*. Invited paper in UNECE meeting, Ljubljana, Slovenia.

Statistics Canada (2007): *Functional Description of the Banff System for Edit and Imputation*, Ottawa.

Vale, S. (2011): *The Generic Statistical Business Process Model and Statistical Data Editing*. Invited paper in UNECE meeting, Ljubljana, Slovenia.

Zhang, L-C. (2011): *Introduction and Presentation from the Network for Industrialization of Editing*. Invited paper in UNECE meeting, Ljubljana, Slovenia.

Teaching Survey Statistics by Teleteaching: A joint project at three German universities

Ulrich Rendtel[1], Florian Meinfelder[2]

[1]Free University of Berlin, e-mail: Ulrich.Rendtel@fu-berlin.de
[2]University of Bamberg, e-mail: florian.meinfelder@uni-bamberg.de

Although surveys have become a frequent tool in economics and social sciences there is not adequate education of students in the design and the analysis of surveys. For Germany there is a need to establish a master program with a strong focus on survey statistics. However, there exists no German university with a minimum of 3 teachers that is needed to run successfully a specialized master program.

We present an initiative of three docents at three universities in three different German federal states who run a joint master program in survey statistics. The program started in the winter term 2010/11. The major feature is the accumulation of teaching modules from the different universities which generates an attractive curriculum. The accumulation is reached by what we call ‘Teleteaching’. This is not only a mere video conference. On a second channel the slide presentation is transmitted in a bi-directional fashion. This channel offers the possibility to communicate between all three sites via a tablet functionality. Thus it is possible that students from Trier may ask the docent in Berlin a question by writing or underlining a formula on the joint slide representation. This approach guarantees a bi-directional learning on the audio, video and formula channel. Meanwhile the technical issues work quite well and we solved the technical problems at the beginning.

The administrative differences of the three German States are not so easy to cope with. There are different teaching terms and different examination rules.

The presentation gives an overview of our concept and presents videos and screenshots of the teaching. The difficulties in the realisation of the concept will be presented also. The web page of the program can be found at Master.Surveystatistics.net.

Contributed papers

Combinining samples of HBS and EU-SILC in Estonia

Julia Aru[1]

[1]University of Tartu, Statistics Estonia, e-mail: julia.aru@gmail.com

Abstract

Statistics Estonia conducts several social surveys with very similar or identical target populations. These surveys focus on different topics, but still contain a block of common questions. Thanks to that, it is possible to combine samples of different surveys to produce more detailed output on these common questions and increase precision. We discuss combining of the samples of two surveys: Survey on Income and Living Conditions (EU-SILC) and Household Budget Survey (HBS). First is a longitudinal survey with rotational design, and second is a purely cross-sectional survey, so the main challenge is computing weights for the combined sample. We discuss several approaches to weighting, gain in precision for combined sample, and final recommendation of the weighting method.

Keywords: Workshop, template, contributed paper

1 Introduction

Statistics Estonia, like any other national statistical office, conducts a lot of social surveys. These surveys focus on various topics and may differ in terms of target population and design, but there is always some overlap in questionnaires, e.g. education, socio-economic status, living conditions etc. In this situation, estimates for common questions can be derived from several surveys. By combining samples of several surveys and estimating common questions from the bigger combined sample, NSI can avoid publishing different estimates for the same phenomenon, which is confusing for users, as well as increase precision of output. This approach is already used for years in several European countries, like Netherlands and UK, while this article will describe the results of the first exercise of this kind in Statistics Estonia.

Two surveys are used in this analysis: Household Budget Survey (HBS) and Survey on Income and Living Conditions (EU-SILC). Features of these surveys are summarized in Table 1.

Table 1: HBS and EU-SILC

Feature	HBS	EU-SILC
Target population	All persons in private households	
Longitudinal component	No longitudinal component, purely cross-sectional	Household remains in the sample for 4 years; a sample of any single year consists of a new part, which is approached for the first time, and repeated part, which is

		approached for the second, third or fourth time.
Sampling design	Stratified systematic sampling of persons from the Population Register, with whole household included along with selected person, which results in PPS sampling for households;	
Non-response correction	Logistic regression with age and gender of selected person, county group and degree of urbanisation.	Logistic regression for new and repeated part separately. Predictors in new part: age and gender of selected person, county group and degree of urbanisation. Predictors in repeated part: tenure status, type of household, county group, nationality, degree of urbanisation, income decile in previous year.
Calibration	Gender-age group and county	
Sample size in 2010	3600 hhs	5000 hhs

The main challenge in this context is computation of weights for combined sample. There are survey-specific weights, which account for the design of the survey, are corrected for non-response and calibrated. Simple method of weighting uses these existing weights after adjusting with a scaling factor as will be described below. More complicated method in some sense starts from the beginning and calculates the probabilities to be included into the combined sample for each household. In the following section we will describe each method in more detail.

2 Methods for weighting the combined sample

2.1 Method of adjusting existing weights

The following method of calculating weights uses existing survey-specific weights and thus is quite simple to implement as the only thing an analyst needs to calculate is an adjustment factor. This method is also referred to as the method on combining samples by Iachan *et al.* (2003) and O'Muircheartaigh & Pedlow (2002). It is simpler to explain to users, more transparent and less dependent on models and assumptions. Nevertheless, it is not clear how applicable it is in case of differences of target populations between the surveys, as will be discussed later.

In general, when adjusting existing weights, we need to calculate a factor $\alpha \in [0,1]$ by which we multiply the weights of the first survey. The weights for the second survey are then multiplied by the factor $1-\alpha$ to ensure that weights for the whole combined sample still sum up the population size. A variety of methods has been proposed for calculating α (see, for example Korn & Graubard, 1999). We use the method described in O'Muircheartaigh & Pedlow (2002), which exploits samples sizes and variability of the survey-specific weights:

$$\alpha = \frac{n_1 / d_1}{n_1 / d_1 + n_2 / d_2}, \text{ where } d_1 = 1 + \frac{Var(w_i, i \in SILC)}{\overline{w}_{SILC}{}^2}, \; d_2 = 1 + \frac{Var(w_i, i \in HBS)}{\overline{w}_{HBS}{}^2},$$

and n_1 and n_2 are respectively EU-SILC and HBS sample sizes.

For the two surveys used in this analysis, $\alpha = 0.561$. Quantities d_1 and d_2 are well-known expressions for the design effect (the part of that due to the variability of weights), and thus n_1/d_1 and n_2/d_2 are effective sample sizes of the surveys. The survey with larger effective sample size receives bigger factor and thus is more influential on the estimates.

Finally, weights were calibrated by 5-year gender-age groups and county.

2.2 Method of cumulating probabilities

Another approach to weighting is to calculate the probability to be included (and respond) in the combined sample. Here, for each household, we calculate the probability to be included in the combined sample as a whole, i.e. to be included in one of the samples, independently on which it was really included in. So, for example, for a household from HBS we need to calculate the probability that it would have been included in EU-SILC, and vice versa, taking into account survey-specific response pattern.

Because of different response models used for different parts of EU-SILC, this survey is divided into the new part and repeated part. In what follows we treat the combined sample as the concatenation of three (instead of two) surveys: EU-SILC repeated part, EU-SILC new part, HBS.

For this method to be comparable with simple method we use the same non-response adjustment methods as described in Table 1. For EU-SILC repeated part, probability to respond in 2010 is modelled as the product of probability to respond in the first year (in the year of first selection into the sample) and probability to respond in 2010 (given the household has responded in the year of selection).

We use following notation:

S – combined sample;

R – response set for the combined sample;

S_1, R_1, R'_1 – respectively the sample, first year response set and 2010 response set for EU-SILC repeated part

S_2, R_2 – sample and 2010 response set for EU-SILC new part;

S_3, R_3 – sample and 2010 response set, HBS;

As $R = R'_1 \cup R_2 \cup R_3$ and surveys are negatively coordinated (households already participating in one of the surveys are dropped prior to the sample selection for the other) the probability of household i to be included into the combined sample and respond can be written as:

$$\begin{aligned} \Pr(i \in R) &= \Pr(i \in R'_1) + \Pr(i \in R_2) + \Pr(i \in R_3) = \\ &= \Pr(i \in R'_1 \mid i \in R_1)\Pr(i \in R_1 \mid i \in S_1)\Pr(i \in S_1) + \\ &+ \Pr(i \in R_2 \mid i \in S_2)\Pr(i \in S_2) + \Pr(i \in R_3 \mid i \in S_3)\Pr(i \in S_3). \end{aligned} \tag{1}$$

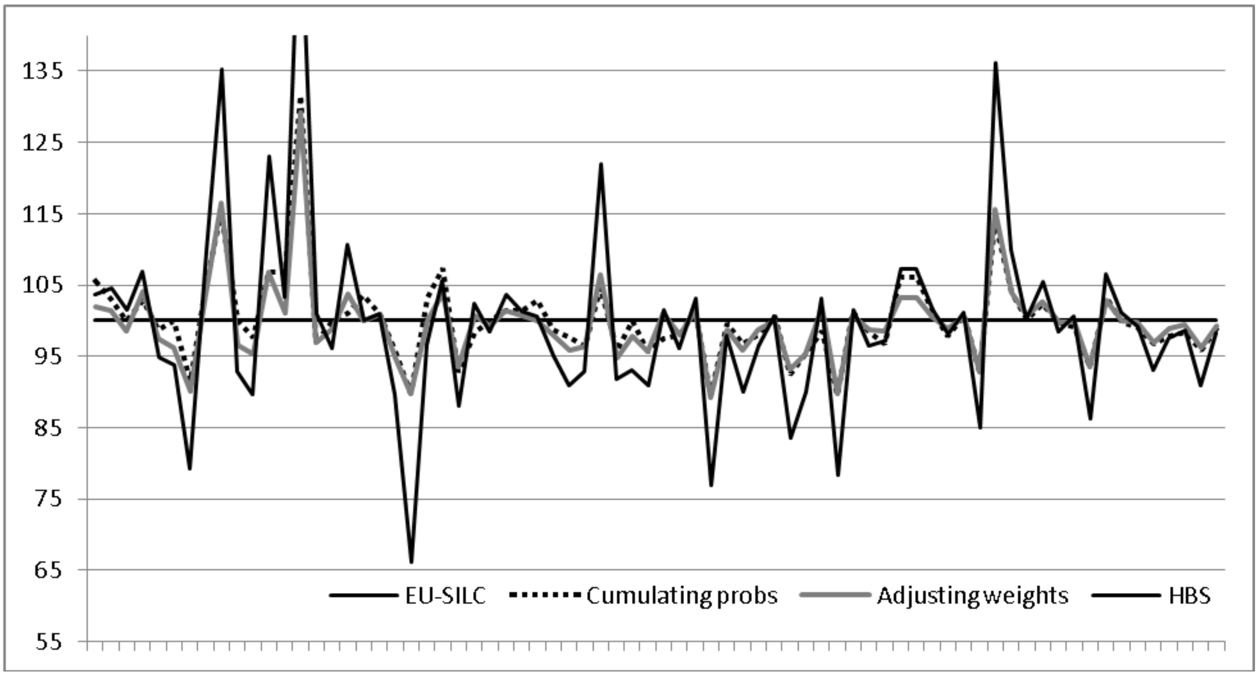
That is, for example, for every household in HBS we need to calculated the inclusion probability and response probability *as if* it is included in EU-SILC new part and *as if* it is included in EU-SILC repeated part. To calculate response probabilities, three response models had to be fitted to the survey-specific data (logistic regression) as shown in Table 1. Fitted logistic regression equation was then applied to every household to calculate response probability, irrespective of the survey the household originated from.

Preliminary weight is a reciprocal of probability (1), and before use it was calibrated by 5-year gender-age groups and county.

3 Comparison of estimates

In spite of different weighting methods for the combined sample, estimates of variables measured in both surveys were very similar. Figure 1 shows a number of estimates calculated with each of the two methods described above as well as the same estimates from specific surveys. All estimates are calculated relative to the EU-SILC estimate, taken as 100%.

Figure 1: Estimates of common variables calculated with different weighting methods



We calculated variance estimates and coeficients of variation for common variables with each of methods, and compared estimates of some parameters with the true values from registers to assess bias. Results are shown in Table 2.

Table 2: Comparison of estimates

	EU-SILC	Cumulating probabilities	Adjusting weights	HBS
Design effect	1,69	1,54	1,66	1,57
Sample size	4972	8604	8604	3632
Effective sample size	2947	5569	5180	2306
Average cv of estimates (%)	3,10	2,28	2,35	3,56
Average absolute relative bias of estimates (%)	17,2	13,5	13,5	21,4

Method of cumulating probabilities seems to give less variable weights, which gives some gain in precision of estimates as compared to the method of adjusting weights. But both methods decreased bias equally and gain in precision appears to be marginal.

4 Summary and future plans

The first exercise on combining sample of two surveys gave very promising results. We compared two methods of weighting: adjusting existing weights and cumulation probabilties. Methods gave very similar results both in terms of precision and bias. So, at least for the two surveys examined, we can use a simpler method of adjusting existing weights. Method of cumulating probabilities is much more difficult to implement, it requires calculating design inclusion probabilities from the beginning and fitting of several response models on different sets of data. For the two surveys at hand, it gave minor gain in precison as compared to the other method, so we suppose it is not worth the effort in future. Statistics Estonia is planning to use the combined sample for producing regular statistics from 2013, with method of adjusting weights as we recommended.

Still, we are going to continue reseach on this topic. It is not clear, would simpler method perform so well in the case of more serious differences between the designs and target populations of the surveys involved. Next step would be to add the Labour Fource Survey (LFS) to the two surveys used in this analysis. LFS has somewhat different design and target population and we are going to repeat this comparison of weighting methods and give recommendations on the weighting methods for combination of three surveys: EU-SILC, HBS and LFS.

Another challenging topic is to re-calibrate survey-spesific weights to the estimates of combined sample. This is also appealing since we don’t know much about first year non-responders in our surveys (available information is limited to register variables such as place of residence, age and gender). For example, with combined sample we could estimate the distribution by education status more precisely (education is considered to be a good predictor for many other topic variables) and then re-calibrate each survey by education. This is expected to improve precision of survey-specific estimates, but till then remains a topic for future research.

References

Iachan, R., Saaverda, P. & Robb, W. (2003) Two weighting schemes for combining sample in the Health Behaviors in School-age Children Survey. *2003 Joint Statistical Meetings - Section on Survey Research Methods*

Korn, E. L. & B. I. Graubard (1999). *Analysis of Health Surveys*. Wiley, New York.

O'Muircheartaigh, C. & Pedlow, S. (2002). Combining samples vs. cumulating cases: a comparison of two weighting strategies in NLSY97. *ASA Proceedings of the Joint Statistical Meetings*, pp. 2557-2562.

The problems of working out of tools household sample surveys

Natalia Bandarenka[1]

[1] Belarus State Economic University, Department of Statistics, e-mail: bondnata@mail.ru

Abstract

The paper considers the problem of developing tools for sample surveys conducted in the Republic of Belarus. The author has carried out a comparative analysis of the main tools with selection of their deficiencies and dignities.

Keywords: Tools, questionnaire, respondents, non-response.

1 Introduction

The active social policy is impossible without studying various aspects of living standards. Now there are three random surveys carried out in the Republic of Belarus: a survey of incomes and expenditures of the population (from 1995), a survey of private subsidiary plots in rural areas (from 2011), a labor force survey (from 2012).

These sample surveys conducted in accordance with international standards, are representative by the volume and structure and have a specific tool. The method of the survey is «face-to-face» interviewing.

2 The main part

For conducted surveys the following tools with the help of international experts have been developed:

– questionnaires for sample surveys;

– card on the dwelling for surveyed household;

– toolkit for interviewers (enumerators), which helps to fill in questionnaires, and includes such topics as the rules of communication with the public (first contact), the rules of filling in the card on the dwelling and procedure for completing the construction of questionnaires, the rules of the survey, the interviewer's personal safety and etc.

– identity of the interviewer.

The main methodological tool is questionnaire (s), preparation of which is based on the following principles:

– content of the questionnaire should be directed to obtain the information necessary for the development of social policy;

– maximum simplicity of the questionnaire, taking into account only those questions that can be obtained from non-response;

– simplicity of the questions' text.

The most worked out are the scheme and the questionnaire of the survey of incomes and expenditures of the households. In the course of the survey such information is collected: demographic characteristics, housing conditions, property security, costs and income of households, etc. Two questionnaires are used: a questionnaire for the main interview and a quarterly questionnaire of the income and expenditure. The main questionnaire consists of two sets of questions: housing and general characteristics of household members (education, social status, employment, etc.). The quarterly questionnaire reflects the value of the total family income and HH's expenditure for durable goods, the cost of market services, housing, health, education, etc. Particular emphasis is placed on the structure of goods and services consumption. The cost of food and nonfood products, indicating their place of purchase, are recorded in most detail.

The information obtained is mainly shown as mean values of the feature: per one household, per member of the household. Total values of indicators are not calculated. Questions on employment and production for private subsidiary plots do not always ensure the representativeness of the data, for example for the region. To obtain such information special surveys of private subsidiary plots in rural areas and labor force surveys are carried out.

For the examination of private subsidiary plots in rural areas, the following questionnaires are used:

- Basic questionnaire including general characteristics of household and private subsidiary plot;

- A questionnaire on the sown areas of crops;

- A questionnaire about the presence and movement of livestock and poultry;

- A diary account of livestock production and feed consumption.

These questionnaires allow you to study in detail the level of rural population development in the republic: the availability of land, acreage, cattle, and amounts of crops and livestock.

At the present time a new source of getting current information on the labor market is being implemented in the statistical practice of the Republic of Belarus – a sample survey of households in order to study the problems of unemployment. The main objectives of the survey are:

• Studying of the labor market status and dynamics;

• Getting the most complete and objective summary statistical data (information) on the number of economically active population, employed, unemployed and economically inactive persons in accordance with the criteria of the International Labour Organization;

• Formation of summary statistical data (information) about the causes of unemployment, methods and duration of job search intention and readiness for employment, as well as the composition of the employed, unemployed and economically inactive population by gender, age, educational level, professional qualification structure.

In line with the objectives the survey program includes questions on the following sections:

A. Information about the respondents: gender, age, marital status, citizenship, education, profession or specialty, number of household members, their relationship to HH;

B. Current economic activity (work): the presence of paid work or gainful employment, the classification of employment status, presence of permanent or temporary work;

C. The first (main) job: type of economic activity, occupation (profession, position), the classification of employment status, presence or absence of employees, and the actual normal work week, reasons for working less than normal working hours or temporary absence;

D. Information about additional (second) work: the presence of additional work in the surveyed week or previous month, type of economic activity, occupation (profession), position, classification by employment status, self-employed, the number of hours actually worked on the additional work.

E. Past activities of persons who are not employed in the surveyed week: the presence of a vacancy, the type of activity and occupation of last job, duration of unemployment, reasons for leaving the last place of employment, the availability of a specialty for unemployed which do not have work experience;

F. Information about the economically inactive persons: social status, reasons for rejection of the job search, the reasons for unwillingness to get started;

G. Employment of production of goods or services in the household: employment in the surveyed week, the production of goods in personal subsidiary plots (PSP), the main type of production and time spent in the surveyed week for manufacturing of industrial goods or services for sale, the main type of manufactured industrial goods (services) in HH and the time spent on their production;

All the questions of the survey program are focused on a form of response: digital, alternative ("yes" or "no"), multivariate, when the answer is chosen from several options being proposed. So, the question of age is given in digital form, the presence of job – in the alternative, the methods of job search response is selected from the proposed options (prompts).

In the development, testing and implementation of sample surveys instruments the presence of a number of problems has been revealed (they are common to all surveys):

- A high level of non-response (25-30%);

- Partial non-response, that is, failures on some issues;

- Underestimation (overestimation) of some indicators of the questionnaire (eg, income, level of accumulated savings) by respondents;

- Obtaining error information for the previous period, or relating to misunderstanding of proposed categorical apparatus (type of activity, status).

Concluding remarks

In households sample surveys conducted in the Republic of Belarus the main tool is the questionnaires. In their design, testing and use a number of common problems appears, such as the need to simplify the structure and scope of the questionnaire, the need for latent problems; the problem is the presence of non-response, including partial non-response.

Possible solutions to these problems – the use of international experience, adjustment of the questionnaires content basing on the results of previous surveys. Another way to solve the issues may be associated with the union of several specialized studies into a single multi-purpose, which, however, will require fundamental changes in statistical practices in the Republic of Belarus.

References

Bandarenka N. (2012). *Working out of methodological recommendations of Labor Force Surveys in the Republic of Belarus (report of national expert)*. Minsk.

Bandarenka N. (2012).The problems of labor market statistics in the Republic of Belarus. – Moskow, 2012: Questions of Statistics, 12, pp. 10-13.

Using of Sample Survey for Building Social Portrait of Alcohol-dependent Persons

Anastacia Bobrova[1]

[1]Institute of Economic of National Academy of Sciences of Belarus, e-mail: nastassiabobrova@mail.ru

Abstract

There is a lot of death among working age population in Belarus. One of the main causes is the high level of alcohol consumption. It is possible to get official data on retail trade of alcohol, but it is difficult to know the main characteristics of person that drinks a lot. There are the main results of 2 sampling in this paper. The differences between real portrait of alcohol-dependent person and his portrait that usual population seems are given.

Keywords Alcohol consumption, social portrait, alcoholic sampling

The problem of alcoholism in Belarus is up to date. Alcohol abuse causes serious damage to man and society. Alcoholism is called problem number 3, after cardiovascular diseases and cancer. The probability of accidents and injuries is increased in people who are intoxicated. Working capacity is reducing, work discipline is deteriorating. Drunk drivers and pedestrians are responsible for a large number of road accidents.

Drunkenness and alcoholism are of interest to researchers in different fields, sociologists, economists, doctors, etc. Of particular importance are the results of calculations, surveys and experiments to determine the socio-economic policy. If significant harm of this phenomenon, the question dealt with alcohol abuse are not fully utilized. This abstract focuses on building the social portrait of alcoholic with the main social and economic features.

In fact, the main source of information on household spending, including for alcohol products in Belarus are the results of a sample survey of households. But the indicators of alcohol consumption derived from officially conducted quarterly sample survey of households are not representative because of its specificity. The main problems are the following:

- Insufficient sample size

- Unavailability of the information on individual categories of the population;

- Misrepresentation of the true costs due to socio-cultural reasons (feeling of awkwardness, embarrassment, a desire to leave the answer arises from the survey participants for fear of dropping your credibility, lose respect, be ridiculed, be condemned, etc.). So for the development of alcohol policies is important to conduct special surveys, allowing a detailed look at the situation with alcoholism in the country.

Special one-time survey conducted in RSPC "Mental Health" in July 2011 were examined 8% of the total number of patients were at the time of examination for medical treatment. Sample - repetition-free, quasi-random. Preference for self-random sample was given by the fact that this type of sample selected in strict accordance with the theory of probability and reflects the variability of trait in the general population. (Bokun, Chernysheva, 1997). Observed proportion of women to men is one to seven. The general population - patients with alcoholism under treatment. The sample had the following characteristics:

the degree of selection – single-stage;

in size - big;

the nature of data collection – single-phase.

To check the representativeness of the sample sampling error by the formula of the fraction corresponding to the unrepeated selection was calculated. (Bokun, Chernysheva, 1997. p.27).

$$\mu_w = \sqrt{\frac{W(1-w)}{n}\left(1-\frac{n}{N}\right)} \qquad (1)$$

μ - sampling error;

w – fraction (women among alcoholism treatment persons);

n – sample frame;

N – general population.

The results of calculation of the amount of sampling error of 5% suggest adequate representation.

The survey included three sets of questions. The first unit is devoted to the socio-economic characteristics of the respondent, in particular about his education level, marital status, changes the level of welfare. The second block contains questions to assess the health status of the respondent. It also asked about smoking. The last section includes questions about the duration of the abuse of alcohol, the types and amounts of alcohol consumed. In addition, the third block contains questions about the causes and consequences of hazardous drinking.

As a result of the sample survey in which respondents had the alcohol desease, further examination was conducted among the entire population of the same sample size and the proportion of men and women. Although sample it is not representative of the whole population, it is interested in the results. Its essence was that the respondents answered the same questions, but in terms of how they seem the alcoholic. Statement respondents answered three additional questions: gender, age and occupation.

Sampling of alcoholism treatment persons

Thus, the average age of alcoholism treatment person was 33 years, 32 for men and 40 for women. The remaining sample results are mainly presented for the total group of respondents, as no differences according to sex. Almost all respondents had secondary specialized vocational education. In one-third of respondents were either divorced or never married. Only 30% were married and had children. Most of the respondents are working with low income (money earned enough only for basic necessities). Only 12% of respondents could afford to purchase home appliances.

70% of respondents believed that their health status was satisfactory. The main problems were a headache and a bad mood, and depression. Among men more than 80% were smoking more than half a pack per day. For women was the reverse situation – 80% did not smoke. More than 30% of the respondents abused alcohol more than 10 years, 25% - about 5 years, the rest – much less. The most common type of alcohol is vodka for men and wine for women. Contrary to the belief that consumed a large number of non-beverage alcohol, the results of a sample survey shows that only 30% of respondents have tried it and only 1% consumed it periodically. And as a non-beverage alcohol was given moonshine, which is consumed in rural areas (as noted by respondents). The results showed that the frequency and amount of alcohol consumed and the percussion is really excessive. Thus, the vast majority of the respondents consumed alcohol more than once a week. For one drink of wine is drunk about 1 bottle (750 ml) bottle of vodka for more than half (300 ml or more).

The respondents believed that the stress of family and personal experiences had pushed them to hazardous drinking. However, 25% of persons showed as a push factor the disability to spend free time in any other way. Among the major consequences - the deterioration of health status (70%), loss of respect from others (40%), family breakdown (30%), poverty (30%).

In conclusion, it should be noted that 80% of alcoholism treatment persons blamed only himself of his weakness, and the rest did not answer the question.

Comparison of results of two surveys

According to respondents, an alcoholic living in predominantly rural areas (65%), while among the alcoholism treatment persons was only 1% from the village. The average age of the patient's alcoholism almost coincided and was 34 years old.

It is important to note that the population is inclined to regard alcoholics worse than they really are. Thus, the majority believed that the level of education average and below the base that he is divorced and has a very low abundance (not even enough for bare necessities.) A similar situation exists in the ratio of the two other blocks. Most respondents indicated that alcoholics have cardiovascular disease and digestive problems. Almost all said that alcoholism treatment persons people smoke. But it was believed that they smoke half a pack of cigarettes.

Very interesting results were obtained from the block on the abuse of alcohol. Among the population there was a perception that all alcoholics was not only tried but also consume non-potable alcohol. In addition, among beverages, beer was indicated rarely. The consumption of samogon and spirits dominated the vodka, respectively, 55 and 45%. The survey showed that the population refered to the alcoholic who drinks every day, and shock doses.

The reason for the hazardous drinking was also different - nearly 60% of respondents believed that alcoholics are those, because the habit had become a dependency, and only 30% attributed this to family and personal problems. According to the survey, the main effects include the following: loss of respect from others (75%), poverty (60%) and family breakdown (50%).

Thus, the survey has allowed us to obtain a social portrait of the alcoholic treatment person in Belarus. This is a man at age 32, who has specialized secondary or vocational education, divorced or unmarried, a lot of smoke. He prefers to drink vodka and low quality cheap wine by shock doses and makes it more than once a week. Comparing the results with the results of the survey population showed some similarities. However, there were large discrepancies. People tend to refer patients to alcoholism only those persons who abuse alcohol every day and those who drink nondrinking alcohol. To some extent this is due to lack of awareness of the depth of the problem of alcoholism and alcohol abuse, i.e., that alcoholic would be a man who drinks low alcohol drink a few times a month.

The results may be useful in substantiating the alcohol policy in terms of education about the dangers of alcohol. At the organization of regular ad hoc surveys may be a kind of monitoring of alcoholic treatment persons.

References

Bokun N., Chernysheva T. (1997) Metody vyborochnyh obsledovanii (The methods of sampling). NII statistiki. Minsk, 1997. - 416 p.

Labour Force Survey in Belarus: determination of sample size, sample design, statistical weighting

Natallia Bokun[1]

[1]Belarus State Economic University, e-mail: nataliabokun@rambler.ru

Abstract

The first experience of forming sampling frames in Belarus for the Labour Force Survey (LFS) is analyzed. Various options for determining the sample size are shown. Some issues of sample design and estimation are considered.

Keywords: sample size, selection, three-stage sample, iterative weighting.

1 Introduction

In Belarus, until recently, data on the size and structure of the labor force have been formed once a year, when calculating the balance of labour resources. The major sources of information on the labour market were as follows: continuous reporting of organizations, administrative sources and census. Despite a rather adequate and detailed measurement of indicators of the economically active population, the existing system of the current account had no possibility of monthly and annual estimates of the actual level of unemployment, which according to the Census 2009, 6-7 times higher than its' recorded amount; did not allow to estimate employment by age, professional groups, to determine the status of employment, underemployment, etc. These factors have caused need for a specialized Labour Force Survey.

Nowadays, the National Statistical Committee of the Republic of Belarus together with some foreign and national experts makes the preparatory work on implementation of the Labour Force Survey (LFS). In November 2011 a test sample survey was conducted. Since 2012 LFS is provided on a regular basis.

The purposes are:

- to obtain empirical statistics on the labour force, economically active population, employed, unemployed;
- to obtain empirical statistics on labour force, employed, unemployed by sex, regions, rural, urban;
- to determine real labour force demand and supply.
- Frequency of the results: quarterly and annual.

LFS data will be widely used for the labour market analysis, assess the actual level of unemployment, making optimal management decisions in the field of employment.

The survey covers the whole country: urban and rural areas in each region. Private households are surveyed. Participation in the survey is voluntary.

The target population comprises all residents aged 15-74 years.

2 Sample size

Calculation of sampling frame, on which representativeness, duration and cost of the survey largely depends, is the most important stage of sampling.

To calculate the *sample size*, with the usage of the appropriate formula, recommended strategy for calculation the sample size is to take into account several factors, connected with sample precision, design-effect (deff), household size, non-responses. These factors are:

- the need to select a key indicator by which the sample size is calculated;
- the precision, needed relative sample error;
- desired confidence level;
- estimated (or known) proportion of the population in the specified target group;
- predicted coverage rate, or prevalence for the specified indicator;
- sample deff;
- average household size;
- adjustment for potential loss of sampled households due to non-response.

As a *key indicator* it is recommended to select one of the most important indicators for the survey and on its basis to estimate the maximum size of sampling frame, yielding an estimate for the minimal (not less than 2.5%) stratum of the population.

Selection of the target group and key indicator includes the following stages:

1. Selection of two or three target populations that comprise small percentages of the total population (1-year, 2-year, 5-year age groups) (Multiple Indicator Cluster Survey Manual (2009), p. 4.8).
2. Review of important indicator based on these groups, ignoring indicators that have very low (less than 5%) or very high (more than 50%) prevalence.
3. Maximal indicator value, calculated for target group (10-15% of the population) is 15-20%.
4. Do not pick from desirable low coverage indicators an indicator that is already acceptably low.

Key indicator, used in Belorussian LFS, is the real unemployment rate (by the Census results). Target groups are economically active populations (rural, urban, by regions, 5-year groups).

Design-effect (deff) describes the influence of sample structure on the value of selection bias, it is defined as a ratio of sample variances of the actual stratified cluster sample (σ_a^2) and of a simple random sample of the same overall sample size (σ^2). International statistical practice has shown that the optimal value of deff is 1.5 (Multiple Indicator Cluster Survey Manual (2009), p. 4.3-4.8), which may be sometimes high.

The sample size formula is used (Bokun, N., Chernysheva, T (1997), p. 44-53; Multiple Indicator Cluster Survey Manual (2009), p. 4.5-4.8, 4.11):

$$n = \frac{4r(1-r)\cdot f \cdot 1.2}{(0.12r)^2 \cdot p \cdot n_h}, \tag{1}$$

where n – required size for the key indicator; 4 – the factor to achieve 95% level of confidence, t-criteria; r – predicted prevalence for the key indicator; 1.2 – essential factor in order to raise the sample size by 20% for non-response; f – the symbol for deff (1.5); 0.12 – recommended relative sample error (95% level of confidence); p – proportion of the total population upon which the indicator (*r*) is based; n_h – average household size.

Several types of the sample size calculations were executed:

- random selection for rural and urban population for each region;
- random selection for Belarus (for target groups);
- random selection for each region;
- stratified sampling for each region.

In the first variant a small surveyed group is the economically active population, according to the second it is the number of economically active population in a particular age range (15-20, 20-24 or 15-74 years). In the third and fourth variants a key indicator is an unemployment rate for the unit of a total population: the proportion of unemployed in the population aged 15-74 years. In this case, there is no need to use the surveyed small groups in the calculation – to determine the sample size for each area the classic formula of the sampling theory is used used (Bokun, N., Chernysheva, T (1997), p. 27-50, 44-53), adjusted for deff, non-response and the number of persons aged 15-74 years per one HH in average.

The examples of sample size determination are given in Tables 1-3.

Table 1 – Sample size for LFS. Variant 2

Target group	Real unemployment rate		Target group size		Average household size, n_h	Number of persons of age 15-74 on average, falling to one HH, n'_h	Predicted sample size	
	persons	%, *r*	to total population, *p*	to 15-74-year group, $p^{/}$			$n_1 = \frac{4r(1-r)\cdot f \cdot 1.2}{(0.12r)^2 \cdot p \cdot n_h}$	$n_2 = \frac{4r(1-r)\cdot 1.5 \cdot 1.2}{(0.12r)^2 \cdot p' \cdot n'_h}$
Economically active population of age 20-24 (565833 persons)	60627	10.7	5.95	7.5	2.43	1.94	28860	28860
Economicalyl active population of age 15-74 in rural area (1051627 persons)	69346	6.6	11.06	14.0	2.43	1.94	26328	26052

Table 2 – Sample size for LFS. Variant 3

Regions	Population of age 15-74, *N*, persons	Number of unemployment, persons	Proportion unemployed in the population aged 15-74 years, *w*	Number of persons of age 15-74 on average, falling to one HH, n'_h	Sample size, *n* , number of households	
					Relative standard error μ =0.06, relative limited error Δ =0.12, (without *deff*)	Relative standard error μ =0.075, relative limited error Δ =0.15, (with *deff*)
Brest region	1073227	50065	0.047	1.92	3502	3380
Vitebsk region	979845	37108	0.038	1.87	4480	4312
Gomel region	1132928	46840	0.041	1.89	4102	3946
Grodno region	829263	31757	0.038	1.87	4474	4308
Minsk	1513844	56293	0.037	2.06	4191	4043
Minsk region	1113871	37345	0.033	1.94	4997	4811
Mogilev region	868907	38511	0.044	1.97	3651	3513
Total	7511885	297919	0.040	1.94	29397	28313

Table 3 – Sample size for LFS. Variant 4

Regions	Population of age 15-74, *N*, persons		Proportion unemployed in the population aged 15-74 years, *w*		Sample size, *n* , number of households	
	urban	rural	urban	rural	Relative standard error μ =0.06, relative limited error Δ =0.12, (without *deff*)	Relative standard error μ =0.06, relative limited error Δ =0.12, (with *deff*)
Brest region	728125	345102	0,048	0,043	1987	2981
Vitebsk region	727698	252147	0,039	0,035	2828	4242
Gomel region	844646	288282	0,040	0,044	2525	3788
Grodno region	589695	239568	0,041	0,032	2773	4160
Minsk	1513844		0,037		4211	6317
Minsk region	631161	482710	0,034	0,033	2570	3855
Mogilev region	670561	198346	0,044	0,046	2209	3314
Total	5705730	1806155	0,040	0,038	19103	28657

Calculation results by different variants have shown that required annual sample size is 26-29 thousands of households, or in average – 28 thousands. Without taking into account non-responses sample size is 22 thousands. Therefore predicted sample fraction is 0.6%, or 22 000 HH. It is planned to examine 7000 HH on a quarterly basis.

3 Sample design

The territorial three-stage sample is used: primary unit – city or village council; secondary unit – census enumeration district or village (zone); final sampling unit – household.

As a sample frame for each stage of the selection the data sets are used which are built by the Census 2009:

- set of cities in each region (the first stage);
- set of village councils in each region (the first stage);
- census enumeration districts in each selected city (the second stage);

- villages (settlements) in each selected village council (the second stage);
- the household totality in each census enumeration district and village (the third stage).

Annual updating of the lists of enumeration areas and HH is assumed.

At each stage units are selected with systematic sampling with the probability that is proportional to population size or to the number of households. Variables used for the stratification are: administrative districts, urban/rural.

The first stage. Towns, including urban settlements and rural councils are selected. At first the towns, which necessarily have to get into the survey, are defined. A criterion of population size for their selection is calculated from the peak value of the interviewer (40 HH), the coefficient of the sample (k = n / N) and the average household size (according to Census 2009 – 2.43): $S_{\tilde{a}} = 40 \cdot \frac{1}{0,006} \cdot 2,43 = 16200$. Thus, the sample includes all the "large" cities with a population 17 thousand people or more. Urban settlements with a population less than 17 thousand people are selected systematically or randomly within each region. Their number depends on the pre-planned number of interviewers and the proportions of the population in small and medium-sized towns (table 4). There are 78 cities to be surveyed (43 large, 35 small and medium-sized), which represent over 38% of the total number of cities in Belarus.

Table 4 – The composition of sampling frame for LFS

Region	Number of cities	Number of village councils	Number of selected		Number of selected households		
			enumeration areas	settlements in the village councils	urban	rural	total
Brest region	13	13	32	22	2560	1560	4120
Vitebsk region	14	10	34	47	2720	1200	3920
Gomel region	14	10	38	17	3040	1200	4240
Grodno region	11	11	28	36	2240	1320	3560
Minsk	1	-	56	-	4480	-	4480
Minsk region	13	16	28	33	2240	1920	4160
Mogilev region	12	8	32	25	2560	968	3528
Total	78	68	248	180	19840	8160	28000

The second stage. In urban areas, enumeration areas according to census are selected, in rural – settlements according to census or village councils accounting. They are selected either according to a predetermined loading and the number of interviewers, or by a combination of random and systematic selection with probability proportional to population size.

The third stage. In the selected sites in urban areas and settlements in rural areas the lists of residential apartments and housing estates are compiled. From an actualized inventory of housing units HHs in urban and rural areas are randomly selected.

4 Statistical weighting

The methodology of weighting is based on the assignment for each individual unit corresponding statistical weight.

HH weights are calculated as reciprocal of overall sample probabilities:

$$B_i = \frac{1}{p_1 \cdot p_2 \cdot p_3}, \tag{2}$$

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

For the case of non-response an additional array of HH is reserved within not less than 20% of the total sample ($28000 \cdot 0{,}2 \approx 6000$).

Individual's weights are based on iterative weighting (Multiple Indicator Cluster Survey Manual (2009); Metodika provedenia bazovyh obsledovanij naselenija (1997)). It is possible to use one of two ways: a) the a simplified method; b) iterative weighting (2 or more iterations).

A simplified method (SM) assumes the calculation of individual weights based on the size of age groups, separately for rural and urban areas:

$$k_{uij} = \frac{S_{ij}}{S_{bij}}, \tag{3}$$

where S_{ij} - individual weight *i*-th gender-are group in urban (rural) area of *j*-th region; S_{ij} - the size of *i*-th gender-are group in urban (rural) area in total population; S_{bij} - the size of *i*-th gender-are group in urban (rural) area, that has been selected within the region.

Iterative weighting (IW) involves several iterations:

Iteration I:

a) weights are calculated separately by sex, urban and rural areas;

b) the first correction coefficient (k1) is calculated; weighted variables are: region, sex, rural/urban;

c) the second correction coefficient (k2) is calculated; variables are: region, sex, 12 five-years groups.

Individual weights are equal within each region, five-year groups, one kind of a settlement.

Iteration II:

At the second iteration the operations are implemented on the subsequent adjustment of the basic weight and intermediate extrapolated data on the same criteria as for the first iteration.

Final individual weights for each five-year group:

$$K_i = B_b \cdot k_1 \cdot k_2 \cdot k_3, \tag{4}$$

where: $B_b = \frac{S_j}{s_j}$; $k_1 = \frac{S_t}{S_E}$; $k_2 = \frac{S_{jt}}{S_{E2}}$; S_j, s_j – population size in *j*-th sex-age group based on the result of the Census and survey; S_t – population size in *t*-th group by rural (urban), sex (on the Census data); S_E – extrapolated population size in *t*-th group (by Bb); S_{jt} – population size in *jt*-th sex-age rural (urban) group; S_{E2} – extrapolated population size in *jt*-th group (by Bb and k1); k_3 – generic correction coefficient, calculated in the second iteration ($k_3 = k_{31} \cdot k_{32} \cdot ... \cdot k_{3n}$).

Preliminary results of iterative weighting for unemployment rate and employment rate, calculated for Mogilev region (Table 5) have shown that received sample population is representative. Relative errors for the region don't exceed 7-8%: for the number of unemployed – 6%, number of employed – 1.8%, unemployment rate – 6.6%.

Table 5 – Indicators of sample representativeness . Mogilev region

Indicators	Characteristic value			Error			
	extrapolated, $Э_x$		in the general population, x	in absolute terms, $\Delta a = \lvert x - Э_x \rvert$		in % $\Delta = \frac{\lvert x - Э_x \rvert}{x}$	
	SM	IW		SM	IW	SM	IW
Number of employed, persons	50516	506231	515876	9360	9645	1.81	1.87
Urban area	400763	402333	412962	12199	10629	2.95	2.57
- Male	192868	194658	205508	12640	10850	6.15	5.28
- Female	207894	207675	207454	440	221	0.21	0.11
Rural area	105754	103898	102914	2840	984	2.76	0.96
- Male	57064	55228	55228	1836	0.3	3.32	0.0006
- Female	48690	48670	47686	1003	984	2.10	2.06
Total number of employed, persons							
- Male	249933	249885	260736	10804	10851	4.14	4.16
- Female	256584	256346	255140	1444	1206	0.57	0.47
Number of unemployed, persons	40624	40510	38511	2113	1899	5.49	4.19
Urban area	31995	32094	29332	2663	2762	9.08	9.42
- Male	19876	20046	18381	1495	1665	8.13	9.06
- Female	12120	12049	10951	1169	998	10.67	9.10
Rural area	8629	8416	9179	550	763	5.99	8.31
- Male	6065	5932	6572	507	640	7.72	9.75
- Female	2564	2485	2607	43	122	1.63	4.69
Number of unemployed (persons) among							
- Male	25940	25977	24953	987	1024	3.96	4.10
- Female	14684	14533	13558	1126	975	8.31	7.19

The results of trial calculations and testing of the first version of methodological and software sampling have shown that the main difficulties are associated with the use of different weighting schemes, determining the number of iterations steps, evaluation of structural indicators of employment and unemployment, the presence of atypical employment on the level of primary units (cities, districts).

5 Concluding remarks

The use of three-stage territorial sampling and iterative weighting provides very reliable information over larger number variables of LFS, conducted in Belarus. However, standard errors, calculated by the level of unemployment, the unemployed, in the context of gender-age groups at regional level are rather high (10-12%). Moreover, under a given load and a limited number of interviewers (200), it is not possible on a quarterly basis to question the estimated number of HH - 28000. On the basis of the selected annual array of HH (28000), built by regions, for each quarter, randomly generated four sub-samples are formed (each includes 7000 HH). If the annual array of information makes it possible to obtain a sufficiently representative data at the level of the republic and regions on most indicators (number of employed, unemployed, the economically active population, employment, unemployment, and in the context of all sex-age groups, the urban and rural areas), the quarterly array makes it possible to design and evaluate the indicators with an acceptable degree of accuracy (10-12%) only at the level of the country.

To improve the representativeness by region the indicators of the survey can be formed on the basis of the three samples – the average for three consecutive quarters. It is possible to increase the number of iterations, to use alternative weighting schemes.

References

BOKUN, N., CHERNYSHEVA, T (1997): Metody vyborochnyh obsledovanij. Minsk.

COCHRAN, W (1997): Sampling techniques. John, Willey and sons, inc. New-York.

Multiple Indicator Cluster Survey Manual. Eurostat, 2009.

Metodika provedenia bazovyh obsledovanij naselenija. Kiev, 2008.

Using calibration in a Survey on Transportation of Goods by Road

Juris Breidaks[1]

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

Abstract

The paper is devoted to the quality analysis of the ongoing survey “Transportation and Turnover of Goods by Road” organised by the Central Statistical Bureau of Latvia (CSB). This is a continuous survey. The stratified simple random sampling is used.

The summary estimation of the variance is compared with the summary GREG estimation of the variance in the paper. These estimations are compared by the sample error.

1 Introduction

The Survey on Transport of Goods by Road was initiated in January 1997 as a pilot project organized by *Eurostat* under the *Phare Programme*. It is a continuous survey where information about the vehicles in the sample is obtained through questionnaires mailed to respondents. The target of survey is to obtain the information about transportation of goods by road performed by transport vehicles registered in Latvia. The main variables of interest are tonnes transported, tonne-kilometres performed and kilometres travelled loaded for total goods road transport.

The survey covers transport vehicles that are owned by legal and natural persons and which at the moment of sample formation had undergone technical inspection and could be lawfully used. The data of the Road Traffic Safety Directorate about vehicle registrations and the number of vehicles that had undergone technical inspection reveal, and could be legally used. Special vehicles such as fire-fighting engines, crane lorries, tower cranes, road repair vehicles and other special vehicles were not included in the survey.

Simple random stratified sampling is used. The weekly sample size is 100 vehicles.

2 Stratification

For 2010 stratification is made by capacity, place of registration of vehicles, year of release of the vehicles, status.

Table 1 – Stratification for 2010

Stratum	Capacity and place of registration of vehicles	Year of release of the vehicles	Status of person
3	3,5t<cap. ≤ 5t, Riga(including the district of Riga)	All	Legal
4	3,5t<cap. ≤ 5t, all Latvia without Riga and the district of Riga	All	Legal
5	5t<cap. ≤10t, Riga(including the district of Riga)	2005-2011	Legal
6	5t<cap. ≤ 10t, Riga(including the district of Riga)	1998-2004	Legal
7	5t<cap. ≤ 10t, Riga(including the district of Riga)	1991-1997	Legal
8	5t<cap. ≤ 10t, all Latvia without Riga and the district of Riga	2005-2011	Legal
9	5t<cap. ≤ 10t, all Latvia without Riga and the district of Riga	1998-2004	Legal
10	5t<cap. ≤ 10t, all Latvia without Riga and the district of Riga	1991-1997	Legal
11	cap.>10t, Riga(including the district of Riga)	2005-2011	Legal
12	cap.>10t, Riga(including the district of Riga)	1998-2004	Legal
13	cap.>10t, Riga(including the district of Riga)	1991-1997	Legal
14	cap.>10t, all Latvia without Riga and the district of Riga	2005-2011	Legal
15	cap.>10t, all Latvia without Riga and the district of Riga	1998-2004	Legal
16	cap.>10t, all Latvia without Riga and the district of Riga	1991-1997	Legal
17	the trucks, Riga(including the district of Riga)	2005-2011	Legal
18	the trucks, Riga(including the district of Riga)	1998-2004	Legal
19	the trucks, Riga(including the district of Riga)	1991-1997	Legal
20	the trucks, all Latvia without Riga and the district of Riga	2005-2011	Legal
21	the trucks, all Latvia without Riga and the district of Riga	1998-2004	Legal
22	the trucks, all Latvia without Riga and the district of Riga	1991-1997	Legal
31	3,5t<cap. ≤ 5t, all Latvia	All	Private
32	5t<cap., all Latvia	All	Private
33	the trucks, all Latvia	All	Private

3 The Horvitz – Thomson estimator and variance

Unit design weights are calculated according to the sampling design and inclusion probabilities of units in the sample – $w_{dh} = \frac{N_h}{n_h}$, where N_h is population size of stratum *h* and n_h \ is the sample size in stratum *h*.

Final weight of unit *i* are calculated as $w_h = w_{dh} \cdot \frac{n_h}{n_h^r} = \frac{N_h}{n_h} \cdot \frac{n_h}{n_h^r} = \frac{N_h}{n_h^r}$, where N_h is population size of stratum *h* and n_h is the number of respondents in stratum *h*.

The Horvitz – Thomson estimator (HT) estimator is

$$\hat{Y}_{HT} = \sum_{i=1}^{n^R} y_i w_i$$

and its estimated variance is

$$\hat{V}\left(\hat{Y}\right)=\sum_{h=1}^{H}\frac{1-\frac{n_h^R}{N_h}}{n_h^R}\frac{1}{n_h^R-1}\sum_{i=1}^{n_h^R}\left(w_i y_i-\frac{1}{n_h^R}\sum_{i=1}^{n_h^R}w_i y_i\right)^2$$

where n^R – number of respondents in sample;

H – number of strata in sampling frame;

y_i – value of study variable of unit i; $y=(y_1,\ldots,y_n)$

w_i – final weight of unit i. $w=(w_1,\ldots,w_n)$

4 GREG estimator and variance

Set of responded transport vehicles in each month is assumed to be a sample. New frame in each month is assumed as population of transport vehicles in beginning of the month.

In each month sample was calibrated on the new frame. The number of respondents in each strata and the capacity of vehicles (legal persons - 3,5t<capacity ≤ 5t, 5t < capacity ≤ 10t, capacity>10t, the trucks, private persons – total capacity.) has been used as auxiliary variables.

Package "sampling" of software *R* is used for the calibration, and g-weights are calculated with the help of a function "calib" from this package. Whereas calibration is based on the "raking" method in the function "calib".

Please observe, that using GREG (Generalized Regression) estimator (calibration), the weights are not equivalent in one stratum within.

The GREG estimator is

$$\hat{Y}_{GREG}=\sum_{i=1}^{n^R}y_i w_i g_i$$

and its estimated variance is

$$\hat{V}_{GREG}\left(\hat{Y}\right)=\sum_{h=1}^{H}\frac{1-\frac{n_h^R}{N_h}}{n_h^R}\frac{1}{n_h^R-1}\sum_{i=1}^{n_h^R}\left(w_i g_i e_i-\frac{1}{n_h^R}\sum_{i=1}^{n_h^R}w_i g_i e_i\right)^2$$

where estimated residual is

$$\hat{e}=y-X_s\cdot\left((X_s\cdot w)^T\cdot X_s\right)^{-1}(X_s\cdot w)^T\cdot y$$

5 Results

The variance of estimates was estimated for 15 indicators – total tonnes transported, total tonne-kilometres performed and total kilometres travelled loaded for total goods road transport and for national goods road transport, and for export goods road transport, and for import goods road transport, and for international goods road transport.

Table 2 – The coefficients of variation for estimates of indicators in year 2010

	Quarter							
	1		2		3		4	
	HT	GREG	HT	GREG	HT	GREG	HT	GREG
TONN	8,01	7,94	8,02	7,19	8,96	8,46	9,60	9,22
TKM	4,46	4,46	4,69	4,74	4,30	4,29	4,95	4,98
KML2	3,80	3,72	3,75	3,73	3,50	3,44	3,89	3,82
TO_N	10,18	10,12	9,57	8,47	10,48	9,90	11,38	10,99
TK_N	8,76	8,79	8,43	8,32	7,70	7,68	7,76	7,72
KM_N	7,19	7,05	6,61	6,24	5,92	5,79	5,96	5,76
TO_EXP	10,22	10,02	15,90	18,32	9,00	9,13	10,58	10,41
TK_EXP	7,92	7,85	7,64	7,68	8,10	8,17	9,22	9,44
KM_EXP	7,52	7,37	6,87	6,84	7,22	7,19	8,53	8,60
TO_IMP	15,95	16,59	18,19	19,01	18,10	17,55	12,38	12,21
TK_IMP	13,43	13,38	12,06	12,12	12,01	11,87	13,08	12,99
KM_IMP	9,14	9,04	8,17	8,20	8,13	7,99	8,32	8,27
TO_INT	11,35	11,25	12,70	12,37	12,91	12,85	12,95	12,90
TK_INT	10,54	10,54	11,18	11,37	10,49	10,40	11,90	11,97
KM_INT	9,37	9,39	9,85	10,07	9,62	9,57	10,39	10,35

Notations for this table

TONN	Tonnes transported for total goods road transport
TKM	Tonne-kilometres performed for total goods road transport
KML2	Kilometres travelled loaded for total goods road transport
TO_N	Tonnes transported for national goods road transport
TK_N	Tonne-kilometres performed for national goods road transport
KM_N	Kilometres travelled loaded for national road transport
TO_EXP	Tonnes transported for export goods road transport
TK_EXP	Tonne-kilometres performed for export goods road transport
KM_EXP	Kilometres travelled loaded for export goods road transport
TO_IMP	Tonnes transported for import goods road transport
TK_IMP	Tonne-kilometres performed for import goods road transport
KM_IMP	Kilometres travelled loaded for import national road transport
TO_INT	Tonnes transported for international goods road transport
TK_INT	Tonne-kilometres performed for international goods road transport
KM_INT	Kilometres travelled loaded for total international road transport

References

Särndal C.-E., Sundström S., “Estimation in the presence of nonresponse and frame imperfections, Statistics Sweden”

Särndal C.-E., Sundström S., (2005) Estimation in Surveys with Nonresponse.Wiley, England.

Särndal C.-E., Swensson B., Wretman J. (1992) *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Lapins, J. (1997) Sampling surveys in Latvia: Current situation, problems and future developments. *Statistics in Transition: Journal of the Polish Statistical Association*, 3, 281-292.

Pandutang V, Sukhatme, Balkrishna V. Sukhatme , Shashikala Sukhatme, C. Asok., Sampling Theory of Surveys with Applications, IOWA State University press and Indian Society of Agricultural Statistics, 14-15

Lohr S. L. (1999) *Sampling: Design and Analysis.* Brooks/Cole Publishing Company, Pacific Grove, Calif.

Guillaume OSIER The linearisation approach implemented by EUROSTAT for the first wave of EU-SILC: What could be done from the second wave onwards? Luxembourg Income Study (LIS), Institut National de la Statistique et des Etudes Economiques (STATEC Luxembourg) and Net-SILC2. Paper prepared for the workshop on standard error estimation and other related sampling issues (Eurostat, 29-30 March 2012)

SPSS Inc. (2012) *SPSS® Syntax Reference Guide*

Liberts. M. (2004) *Quality Analysis of a Sample Survey on Transportation of Goods by Road*.

Comparison of Energy Resource Survey Results

Baiba Buceniece[1]

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

Abstract

Annual enterprise survey of the energy resource acquisition and consumption is carried out by the Central Statistical Bureau of Latvia. The aim of the study is comparison of survey results of 2010 and 2011.

In the survey more than 600 variables are collected, but main variables of interest are:

- amount of received heat,
- consumption of electricity,
- consumption of petrol,
- consumption of diesel fuel,
- consumption of natural gas.

Sampling design used for this survey is stratified simple random sample.

The sampling frame is made from Statistical Business Register on November of the reference year. Sampling frame includes economically active merchants, state and municipal budget authorities and agricultural and fish farms with 10 or more employees. Sampling frame for 2010 consists of 60321 statistical units and for 2011 of 63565 statistical units.

The stratification of the sampling frame is based on economic activity (NACE Rev. 2) and on turnover or number of employees.

Sample sizes for 2010 and for 2011 are computed differently. For 2010 sample sizes are calculated using optimized Neyman allocation based on turnover of enterprises (or on number of employees in some stratas, where turnover is missing).

For 2011 the survey data of 2010 are used for calculation of sample size. For each strata five different sample sizes are calculated based on five main variables of interest and the final sample size in each strata is calculated as an average of these five.

Also different weighting procedures are used for 2010 and 2011.

For 2010 several weights are computed:

- weights for variable ”amount of received heat”;
- weights for variable ”consumption of natural gas”;
- weights for variables ”consumption of petrol” and ”consumption of diesel fuel”;
- weights for variable ”consumption of electricity” (weights for this variable are calibrated using auxiliary information, that is available only in aggregated level (not available for each unit (enterprise) separately);

- weights for many variables associated with consumption of wooden fuel;
- weights for other survey variables.

For 2011 only one kind of weights is computed for all survey variables. Weights are calibrated using auxiliary information about delivered electricity and natural gas, that also is available only in aggregated level (not available for each unit (enterprise) separately).

The results of the study - variances and coefficients of variation of main survey variables will be presented in poster session during the workshop.

Count of persons in collective Households in Latvia

Ance Ceriņa[1]

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

Abstract

The population estimates in private and collective households are considered.

Keywords: Collective households, Census 2011

1 Count of person in collective households

There is difference in how a count of persons in collective households was calculated before and after Census 2011.

The count of persons in collective households was calculated using information from different sources before year 2011:

- prisons
- student halls of residence
- army barracks (2005)
- specialized child care centres (boarding schools)
- seniors and persons with mental disorders homes
- night shelters (since 2009)
- children's homes.

The main problem: information came in different time; there were years when previous information from last year was used.

There is huge difference between counts of persons in collective households from 2005 to 2012 (Table 1).

Table 1: Count of person in collective households

	2005	2006	2007	2008	2009	2010	2011	2012
Latvia	**33 970**	**32 186**	**12 734**	**45 967**	**27 082**	**26 871**	**24 754**	**24 754**
Riga	10 587	13 977	2 575	16 583	9 469	9 408	7 340	7 340
Cities	6 597	9 269	1 712	11 333	10 363	10 302	8 707	8 707
Rural area	8 588	8 940	8 447	18 051	7 250	7 161	8 707	8 707

There are two options how to calculate the count of persons in collective households after Census 2011:

- use the count of persons in collective households from Census 2011 and not to update it
- use the count of persons in collective households from Census 2011 and update those sections which are known.

2 Count of person in private households

Total population in Latvia or the count of people in private households is used for calibration of survey data. More precise estimates are obtained when survey data is calibrated using the count of persons in private households.

The count of persons living in private households is a difference between total population of Latvia and count of persons in collective households.

3 Results

The main results: there were only two surveys (Households budget survey and SILC) which were calibrated using count of persons in private households. Other surveys were calibrated using total population of Latvia before Census 2011. There are three more surveys where data are calibrated using count of persons in private households after Census 2011.

The results of the study will be presented in poster session during the workshop.

Accuracy investigation of the composite estimators in the case of sample rotation for two-phase sampling scheme

Chadysas Viktoras[1]

[1]Vilnius Gediminas Technical University, e-mail: viktoras.chadysas@vgtu.lt

Abstract

In this paper we focus on construction of the combined ratio type estimator of the finite population total and their variances in the case of sample rotation for two-phase sampling scheme. We construct composite estimators of the finite population total without and with the use of auxiliary information known from the previous survey. Two types of sampling design are used for sample selection in each of the phases: simple random sampling without replacement and successive sampling (unequal probability sampling without replacement). A simulation study, based on the real population data, is performed and the proposed estimators are compared.

Keywords: Sample rotation, ratio estimator, composite estimator, successive sampling

1 Introduction

The Labour Force Survey (LFS) quarterly provides estimates of the number of employed and unemployed individuals. Repeated sampling from a finite population is a sampling procedure used for the LFS.

Let us consider a finite household population $\mathcal{U} = \{1, \ldots, i, \ldots, N\}$ of size N. For each household, the number of its members is denoted by m_i, $i = 1, 2, \ldots, N$. Then the sum of the household members can be obtained by $M = \sum_{i=1}^{N} m_i$. Let us say that the survey variable y characterizes the number of employed and unemployed individuals in each household. The values y_i of the variable belongs to the set of integers $\{0, 1, \ldots, m_i\}$. We are interested in the estimation of the number of employed (and unemployed) individuals in the population:

$$t_y = \sum_{i=1}^{N} y_i. \tag{1}$$

The previous wave survey data can be used as auxiliary information for estimation of the population total in order to reduce the variance of the estimator. The effectiveness of ratio estimators in the case of any sampling design is described by many authors. If an auxiliary variable is well correlated with the study variable, then it is possible to obtain a more accurate estimate of a parameter.

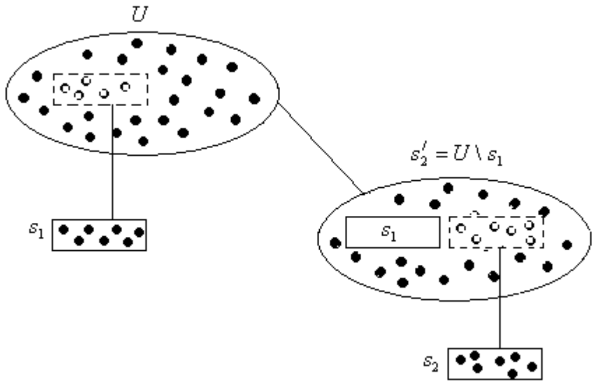
2 Sampling rotation and sample selection

LFS is conducted quarterly at Statistics Lithuania. All members of a household are included into the samples for two subsequent quarters, excluded for the next two quarters, and included into the sample once more for another two quarters.

Two-phase sampling scheme as shown in Fig 1. can be used for the estimation of the population total t_y (1).

The sample rotation procedure for the two-phase sampling scheme presented in Fig 1. shows that the whole sample s consists of two samples: s_1 and s_2. These sample parts are expressed as follows:

Figure 1: The sample selection procedure



$s_1 : \mathcal{U} \longrightarrow s_1;$
$s_2 : \mathcal{U} \longrightarrow s_2' = \mathcal{U} \setminus s_1 \longrightarrow s_2;$ (two phases)

3 Simple estimators of the population total

Firstly, we construct two separate estimators of the total t_y using data of samples s_1 and s_2 respectively. Secondly, we propose a combined estimator of the total using the sample rotation scheme (see section 4).

Step 1

The sample s_1 is selected from the finite population: $\mathcal{U} \longrightarrow s_1$. The corresponding first and second order inclusion probabilities for elements of the sample s_1 are respectively:

$$\pi_{1i} = \mathrm{P}(s_1 \subset \mathcal{U} : i \in s_1);$$
$$\pi_{1ij} = \mathrm{P}(s_1 \subset \mathcal{U} : i \in s_1, j \in s_1).$$

Unbiased Horvitz-Thompson (HT) (Horvitz & Thompson, 1952) estimator of the population total:

$$\hat{t}_{1y}^{HT} = \sum_{i \in s_1} \frac{y_i}{\pi_{1i}}. \tag{2}$$

The variance of the estimator $\hat{t}_{1y}^{HT}$ is

$$Var(\hat{t}_{1y}^{HT}) = \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} (\pi_{1ij} - \pi_{1i}\pi_{1j}) \frac{y_i}{\pi_{1i}} \frac{y_j}{\pi_{1j}}. \tag{3}$$

The variance $Var(\hat{t}_{1y}^{HT})$ is estimated unbiasedly by

$$\widehat{Var}(\hat{t}_{1y}^{HT}) = \sum_{i \in s_1} \sum_{j \in s_1} (1 - \frac{\pi_{1i}\pi_{1j}}{\pi_{1ij}}) \frac{y_i}{\pi_{1i}} \frac{y_j}{\pi_{1j}}. \tag{4}$$

The values of the study variable y in the previous survey can be used as auxiliary information. Let us denote the study variable of the previous wave survey by x with the values x_i and the same variable on the current wave by y with the values y_i, $i \in s_1$. We can form the ratio estimator $\hat{t}_{1y}^{rat}$ of the population total t_y by

$$\hat{t}_{1y}^{rat} = \hat{t}_x^{1w} \frac{\hat{t}_{1y}^{HT}}{\hat{t}_{1x}^{HT}} = \hat{t}_x^{1w} \hat{r}, \tag{5}$$

where

$$\hat{t}_{1x}^{HT} = \sum_{i\in s_1} \frac{x_i}{\pi_{1i}}; \quad \hat{r} = \frac{\hat{t}_{1y}^{HT}}{\hat{t}_{1x}^{HT}}.$$

$\hat{t}_x^{1w}$ – an estimate of the population total t_y calculated using all values of the study variable y in the previous survey, $\hat{t}_{1y}^{HT}$ given in (2). The estimator $\hat{t}_{1y}^{rat}$ is non-linear. Usually the Taylor expansion is used to find its properties.

An approximate variance of the ratio estimator $\hat{t}_{1y}^{rat}$ is:

$$AVar(\hat{t}_{1y}^{rat}) = \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}}(\pi_{1ij} - \pi_{1i}\pi_{1j})\frac{y_i - rx_i}{\pi_{1i}}\frac{y_j - rx_j}{\pi_{1j}}, \tag{6}$$

with $r = \sum_{i\in\mathcal{U}} y_i / \sum_{i\in\mathcal{U}} x_i$.

The variance $AVar(\hat{t}_{1y}^{rat})$ is estimated by

$$\widehat{Var}(\hat{t}_{1y}^{rat}) = \sum_{i\in s_1}\sum_{j\in s_1}(1 - \frac{\pi_{1i}\pi_{1j}}{\pi_{1ij}})\frac{y_i - \hat{r}x_i}{\pi_{1i}}\frac{y_j - \hat{r}x_j}{\pi_{1j}}, \tag{7}$$

using $\hat{r}$ given in (5)

Step 2

The sample s_2 is obtained in two-phase sampling: $\mathcal{U} \longrightarrow s_2' = \mathcal{U} \setminus s_1 \longrightarrow s_2$. The corresponding first and second order inclusion probabilities for samples s_2' (first phase) and s_2 (second phase) are respectively:

$$\begin{aligned}
\pi_{2i}' &= \mathrm{P}(s_2' \subset \mathcal{U} : i \in s_2') = \mathrm{P}(i \notin s_1) = 1 - \mathrm{P}(i \in s_1);\\
\pi_{2ij}' &= \mathrm{P}(s_2' \subset \mathcal{U} : i \in s_2', j \in s_2')\\
&= 1 - \mathrm{P}(i \in s_1, j \in s_1) - \mathrm{P}(i \in s_1, j \notin s_1) - \mathrm{P}(i \notin s_1, j \in s_1);\\
\pi_{2i|s_2'} &= \mathrm{P}(s_2 \subset s_2' : i \in s_2 \mid s_2');\\
\pi_{2ij|s_2'} &= \mathrm{P}(s_2 \subset s_2' : i \in s_2, j \in s_2 \mid s_2').
\end{aligned}$$

Under two-phase sampling, using the π^* estimator (Särndal *et al.*, 1992), the population total t_y is unbiasedly estimated by

$$\hat{t}_{2y}^{(2)} = \sum_{i\in s_2} \frac{y_i}{\pi_{2i}'\pi_{2i|s_2'}}. \tag{8}$$

In the case of two-phase sampling the variance of the estimator $\hat{t}_{2y}^{(2)}$ may be expressed

$$\begin{aligned}
Var(\hat{t}_{2y}^{(2)}) &= \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}}(\pi_{2ij}' - \pi_{2i}'\pi_{2j}')\frac{y_i}{\pi_{2i}'}\frac{y_j}{\pi_{2j}'}\\
&\quad + \mathrm{E}\sum_{i,j\in s_2'}(\pi_{2ij|s_2'} - \pi_{2i|s_2'}\pi_{2j|s_2'})\frac{y_i}{\pi_{2i}'\pi_{2i|s_2'}}\frac{y_j}{\pi_{2j}'\pi_{2j|s_2'}}.
\end{aligned} \tag{9}$$

The variance $Var(\hat{t}_{2y}^{(2)})$ is estimated unbiasedly by

$$\begin{aligned}
\widehat{Var}(\hat{t}_{2y}^{(2)}) &= \sum_{i,j\in s_2}\frac{\pi_{2ij}' - \pi_{2i}'\pi_{2j}'}{\pi_{2ij}'\pi_{2ij|s_2'}}\frac{y_i}{\pi_{2i}'}\frac{y_j}{\pi_{2j}'}\\
&\quad + \sum_{i,j\in s_2}\frac{\pi_{2ij|s_2'} - \pi_{2i|s_2'}\pi_{2j|s_2'}}{\pi_{2ij|s_2'}}\frac{y_i}{\pi_{2i}'\pi_{2i|s_2'}}\frac{y_j}{\pi_{2j}'\pi_{2j|s_2'}}
\end{aligned} \tag{10}$$

4 Combined estimators of the population total for two-phase sampling scheme

In this section the construction of the combined estimators and their variances of the finite population total (1) in the case of sample rotation for two-phase sampling scheme is presented.

By a linear combination of $\hat{t}_{1y}^{HT}$ and $\hat{t}_{2y}^{(2)}$ we obtain an estimator without the use of auxiliary information of the total

$$\hat{t}_y = \frac{1}{2}\hat{t}_{1y}^{HT} + \frac{1}{2}\hat{t}_{2y}^{(2)}. \tag{11}$$

The variance of estimator (11) of the total t_y can be expressed

$$Var(\hat{t}_y) = \Big(\frac{1}{2}\Big)^2 Var(\hat{t}_{1y}^{HT}) + \Big(\frac{1}{2}\Big)^2 Var(\hat{t}_{2y}^{(2)}) + \frac{1}{2}Cov(\hat{t}_{1y}^{HT}, \hat{t}_{2y}^{(2)}). \tag{12}$$

The variance $Var(\hat{t}_2)$ is estimated unbiasedly by

$$\widehat{Var}(\hat{t}_y) = \Big(\frac{1}{2}\Big)^2 \widehat{Var}(\hat{t}_{1y}^{HT}) + \Big(\frac{1}{2}\Big)^2 \widehat{Var}(\hat{t}_{2y}^{(2)}) + \frac{1}{2}\widehat{Cov}(\hat{t}_{1y}^{HT}, \hat{t}_{2y}^{(2)}). \tag{13}$$

By a linear combination of $\hat{t}_{1y}^{rat}$ and $\hat{t}_{2y}^{(2)}$ we obtain an estimator with the use of auxiliary information of the total

$$\hat{t}_y^{rat} = \frac{1}{2}\hat{t}_{1y}^{rat} + \frac{1}{2}\hat{t}_{2y}^{(2)}. \tag{14}$$

The variance of estimator (14) of the total t_y can be expressed

$$AVar(\hat{t}_y^{rat}) = \Big(\frac{1}{2}\Big)^2 AVar(\hat{t}_{1y}^{rat}) + \Big(\frac{1}{2}\Big)^2 Var(\hat{t}_{2y}^{(2)}) + \frac{1}{2}Cov(\hat{t}_{1y}^{rat}, \hat{t}_{2y}^{(2)}). \tag{15}$$

The variance $Var(\hat{t}_2^{rat})$ is estimated by

$$\widehat{Var}(\hat{t}_y^{rat}) = \Big(\frac{1}{2}\Big)^2 \widehat{Var}(\hat{t}_{1y}^{rat}) + \Big(\frac{1}{2}\Big)^2 \widehat{Var}(\hat{t}_{2y}^{(2)}) + \frac{1}{2}\widehat{Cov}(\hat{t}_{1y}^{rat}, \hat{t}_{2y}^{(2)}). \tag{16}$$

Further we are interested in the estimation of the finite population total t_y using two-phase sampling scheme, when simple random samples of households without replacement and samples with probabilities proportional to household size without replacement are drawn at each of the phases. Data of two quarters is used for the estimation of the population total t_y.

4.1 Simple random sampling (SRS) of households without replacement

Assume that s_1 of size n_1 is a simple random sample from the population $\mathcal{U}$ and its complement $s_2' = \mathcal{U}\backslash s_1$ of size $N - n_1$ is also a simple random sample from the population $\mathcal{U}$. s_2 of size n_2 is a simple random sample from s_2'. Then the first and second order inclusion probabilities to be used for (11), (13), (14) and (16) are calculated as follows:

$$\begin{aligned}
\pi_{1i} &= \mathrm{P}(s_1 \subset \mathcal{U} : i \in s_1) = \frac{n_1}{N};\\
\pi_{1ij} &= \mathrm{P}(s_1 \subset \mathcal{U} : i \in s_1, j \in s_1) = \frac{n_1(n_1-1)}{N(N-1)};\\
\pi_{2i}' &= \mathrm{P}(s_2' \subset \mathcal{U} : i \in s_2') = \frac{N-n_1}{N};\\
\pi_{2ij}' &= \mathrm{P}(s_2' \subset \mathcal{U} : i \in s_2', j \in s_2') = 1 - \frac{n_1(n_1-1)}{N(N-1)} - 2\frac{n_1}{N-1}\Big(1 - \frac{n_1}{N}\Big);\\
\pi_{2i|s_2'} &= \mathrm{P}(s_2 \subset s_2' : i \in s_2 \mid s_2') = \frac{n_2}{N-n_1};\\
\pi_{2ij|s_2'} &= \mathrm{P}(s_2 \subset s_2' : i \in s_2, j \in s_2 \mid s_2') = \frac{n_2(n_2-1)}{(N-n_1)(N-n_1-1)}.
\end{aligned}$$

We estimate the variances of the estimators $\hat{t}_y$ and $\hat{t}_y^{rat}$ of the total t_y in (13) and (16) respectively with

$$\widehat{Cov}(\hat{t}_{1y}^{HT}, \hat{t}_{2y}^{(2)}) = -\frac{N}{(N-n_1)(n_1-1)}\sum_{i\in s}(y_i-\bar{y})^2 \tag{17}$$

and

$$\widehat{Cov}(\hat{t}_{1y}^{rat}, \hat{t}_{2y}^{(2)}) = -\frac{N^2}{(N-n_1)(n_1-1)}\sum_{i\in s_1}(y_i-\hat{r}x_i)(y_i-\bar{y}) \tag{18}$$

4.2 Unequal probability sampling of households without replacement with probability proportional to its size (PPS) - successive sampling

Suppose $\mathcal{U} = \{1,\dots,i,\dots,N\}$ is a finite population of size N, here N is the number of the households. Let us assume that m_i is the number of the household members, $M = \sum_{i=1}^{N} m_i$ is the sum of the households members.

Assume that sample s_1 of size n_1, sample $s_2' = \mathcal{U}\setminus s_1$ of size $N-n_1$ and sample s_2 of size n_2 is drawn according to an order sampling design introduced by Rosen (1996) - successive sampling. To each unit i in the population $\mathcal{U}$, the random number U_i having uniform distribution $U(0,1)$ is generated. Let us assume that $\lambda_i = \frac{n_1 m_i}{M}$, then $Q_i = -ln(1-U_i)/-ln(1-\lambda_i)$, $i\in\mathcal{U}$ are the ranking variables. Values of Q_i are found, the population elements with the n_1 smallest Q values constitute the sample s_1. Sample $s_2' = \mathcal{U}\setminus s_1$ of size $N-n_1$ is also obtained by a successive order sampling. Now, to each unit i in sample s_2' the random number U_i having uniform distribution $U(0,1)$ is generated, values of ranking variables $Q_i = -ln(1-U_i)/-ln(1-\lambda_i)$, $i\in s_2'$, are recalculated with $\lambda_i = \frac{n_2 m_i}{\sum_{i\in s_2'} m_i}$, and sample s_2' elements with the n_2 smallest Q values constitute the sample s_2.

Then the first order inclusion probabilities to be used for the estimation of the total t_y in (11) and (14) are calculated as follows:

$$\begin{aligned}\pi_{1i} &= \mathrm{P}(s_1\subset\mathcal{U} : i\in s_1) \approx \lambda_i = \frac{n_1 m_i}{M};\\ \pi_{2i}' &= \mathrm{P}(s_2'\subset\mathcal{U} : i\in s_2') \approx \lambda_{2i}' = \frac{M-n_1 m_i}{M};\\ \pi_{2i|s_2'} &= \mathrm{P}(s_2\subset s_2' : i\in s_2 \mid s_2') \approx \lambda_{2i} = \frac{n_2 m_i}{M_2};\quad M_2=\sum_{i\in s_2'} m_i.\end{aligned}$$

There are expressions for second order inclusion probabilities π_{1ij} in the case of order sampling design presented in (Aires, 1999), but they are too complex and time consuming. Therefore, we approximate π_{1ij} for successive sampling design by $\check{\pi}_{1ij}$ for unequal probability without replacement conditional Poisson sampling design, presented by Aires (1999):

$$\begin{aligned}\check{\pi}_{1ij} &= \frac{1}{\gamma_{1i}-\gamma_{1j}}(\gamma_{1i}\check{\pi}_{1j}-\gamma_{1j}\check{\pi}_{1i}), &&\text{for } \gamma_{1i}\neq\gamma_{1j},\\ \check{\pi}_{1ij} &= \frac{1}{k_{1i}}\left((n_1-1)\check{\pi}_{1i}-\sum_{j:\ \gamma_{1j}\neq\gamma_{1i}}\check{\pi}_{1ij}\right), &&\text{for } \gamma_{1i}=\gamma_{1j},\quad i,j\in\mathcal{U},\quad i\neq j.\end{aligned}$$

Here k_{1i} is the number of elements with $j\neq i$ such that $\gamma_{1i}=\gamma_{1j}$ and $\gamma_{1i}=\frac{p_{1i}}{1-p_{1i}}$. The probability p_{1i} is a selection probability of element i for conditional Poisson sampling design, $i\in\mathcal{U}$. Since p_{1i} are unknown, we use approximation result of Bondesson *et al.* (2006):

$$\frac{p_{1i}}{1-p_{1i}} \propto \frac{\lambda_{1i}}{1-\lambda_{1i}}exp\left(\frac{\frac{1}{2}-\lambda_{1i}}{d}\right),\quad d=\sum_{i=1}^{N}\lambda_{1i}(1-\lambda_{1i}).$$

Let us approximate the first order inclusion probabilities:

$$\pi_{1i} = \breve{\pi}_{1i} \approx \lambda_i = \frac{n_1 m_i}{M};$$
$$\pi'_{2i} = \breve{\pi}'_{2i} \approx \lambda'_{2i} = \frac{M - n_1 m_i}{M};$$
$$\pi_{2i|s'_2} = \breve{\pi}_{2i|s'_2} \approx \lambda_{2i} = \frac{n_2 m_i}{M_2}; \quad M_2 = \sum_{i \in s'_2} m_i.$$

The second order inclusion probabilities for samples s'_2 and s_2 are taken as follows:

$$\begin{aligned}
\pi'_{2ij} &= \mathrm{P}(s'_2 \subset \mathcal{U} : i \in s'_2, j \in s'_2) \\
&\cong 1 - \breve{\pi}_{1ij} - \frac{n_1 m_i}{M - m_j}(1 - \lambda_j) - \frac{n_1 m_j}{M - m_i}(1 - \lambda_i); \\
\breve{\pi}_{2ij|s'_2} &= \frac{1}{\gamma_{2i} - \gamma_{2j}}(\gamma_{2i}\breve{\pi}_{2j|s'_2} - \gamma_{2j}\breve{\pi}_{2i|s'_2}), && \text{for } \gamma_{2i} \neq \gamma_{2j}, \\
\breve{\pi}_{2ij|s'_2} &= \frac{1}{k_{2i}}\left((n_2 - 1)\breve{\pi}_{2i|s'_2} - \sum_{j:\ \gamma_{2j} \neq \gamma_{2i}} \breve{\pi}_{2ij|s'_2}\right), && \text{for } \gamma_{2i} = \gamma_{2j}.
\end{aligned}$$

Here k_{2i} is the number of elements $j \neq i$ such that $\gamma_{2i} = \gamma_{2j}$ and $\gamma_{2i} = \frac{\lambda_{2i}}{1-\lambda_{2i}} exp\left(\frac{\frac{1}{2} - \lambda_{2i}}{d}\right)$, $d = \sum_{i \in s'_2} \lambda_{2i}(1-\lambda_{2i})$. We have $\breve{\pi}_{2ii|s'_2} = \breve{\pi}_{2i|s'_2}$ in the case $i = j$. After replacing the π values in (4), (7) and (10) with the corresponding values presented in this section we estimate the variance of the estimators $\hat{t_2}$ and $\hat{t}_2^{rat}$ of the total t_y in (13) and (16) respectively with

$$\widehat{Cov}(\hat{t}_{1y}^{HT}, \hat{t}_{2y}^{(2)}) = -\frac{M^2}{(N - n_1)n_1} \sum_{i \in s_1} \frac{1}{m_i^2}\left(y_i - \frac{\hat{t}_{1y}^{HT}}{\hat{N}}\right)^2 \tag{19}$$

and

$$\widehat{Cov}(\hat{t}_{1y}^{rat}, \hat{t}_{2y}^{(2)}) = -\frac{M^2}{(N - n_1)n_1} \sum_{i \in s_1} \frac{1}{m_i^2}\left(y_i - \frac{\hat{t}_{1y}^{HT}}{\hat{N}}\right)\left(y_i - \frac{\hat{t}_{1y}^{HT}}{\hat{t}_{1x}^{HT}} x_i\right), \tag{20}$$

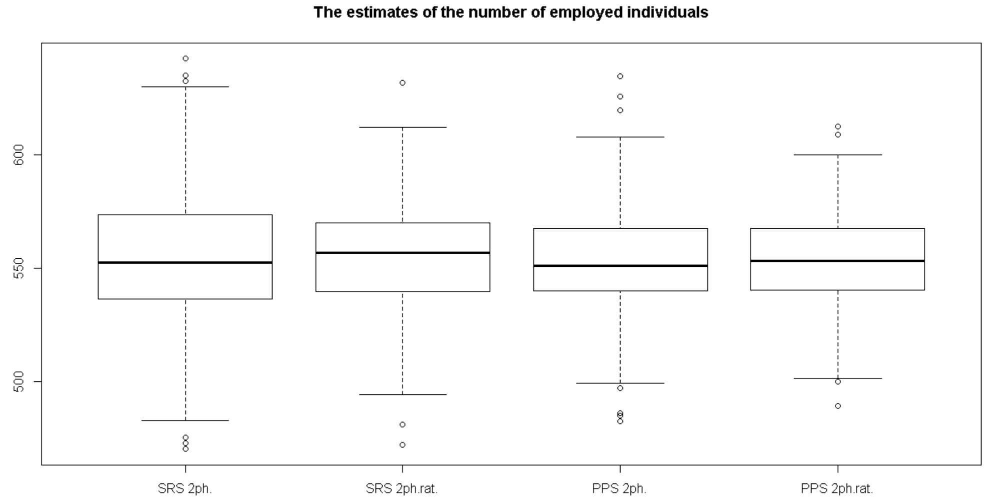
where $\hat{N} = \sum_{i \in s_1} 1/\pi_{1i}$.

5 Simulation study

In this section, we present the simulation study for the comparison of the performance of proposed estimators of the total using data of two quarters, with simple random sampling (SRS) and unequal probability sampling (successive order sampling) (PPS) of households without replacement in each of the phase. We study real LFS data of Statistics Lithuania. The study population consists of N=500 households. The variables of interest, y and x, are the number of employed and unemployed individuals in the population of households. We have selected B=500 samples s_1 and s_2 of size $n_1 = n_2 = 100$ ($n = n_1 + n_2 = 200$) by the simple random sampling and successive sampling scheme.

For each of the estimators $\hat{t}_y$ (11) and $\hat{t}_y^{rat}$ (14), we have calculated the estimates of the population total of the study variable y. Also are calculated the averages of the estimates, averages of the estimates of variances. The results of the simulation are presented in Fig 2. and Fig 3.

Figure 2: Box-Plot diagrams of the estimates of the number of employed and the number of unemployed



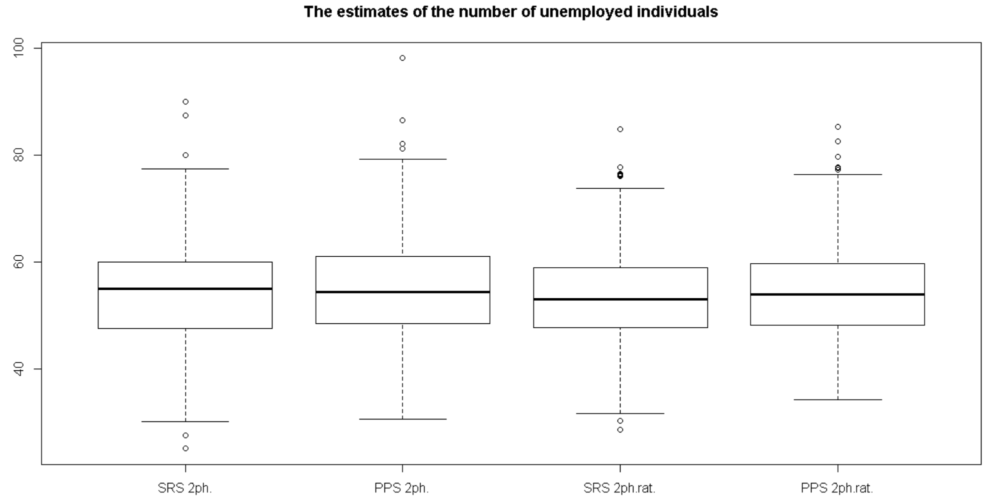
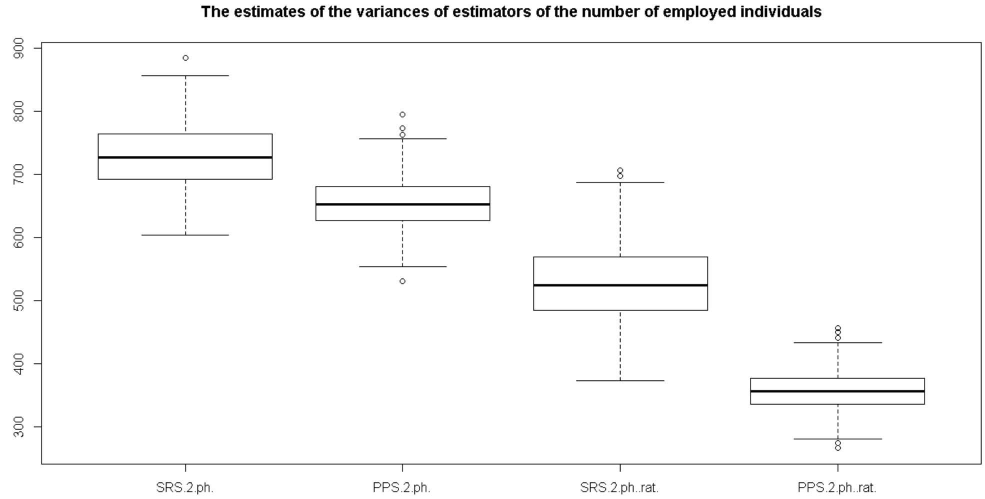
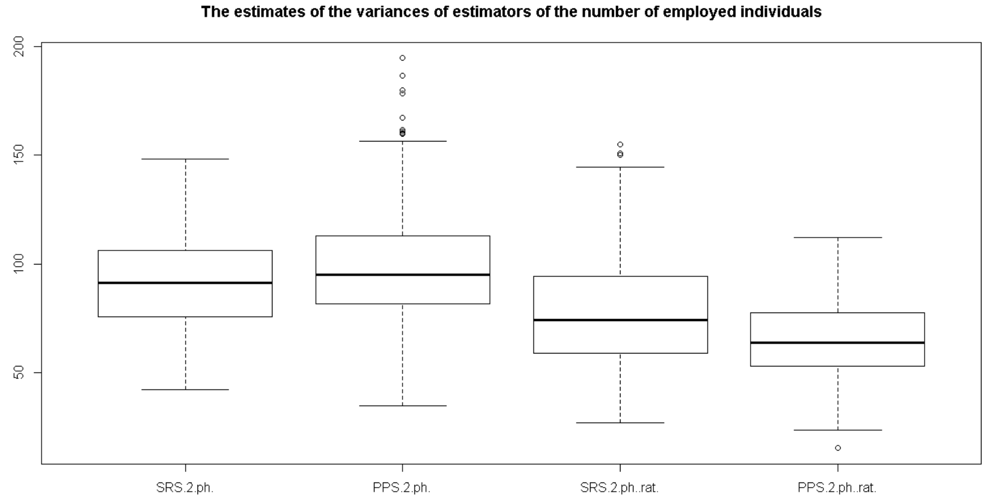


Figure 3: Box-Plot diagrams of estimates of the variances of estimators of the number of employed and the number of unemployed





6 Conclusions

The study addresses a practical problem related to the estimation of the number of employed and unemployed persons in two-phase sampling with simple random sampling and successive sampling design in each of the sampling phases. The successive sampling design is effective to estimate the number of employed and its effectiveness is the same as for SRS to estimate the number of the unemployed in the Lithuanian LFS. The ratio type combined estimator in two-phase LFS sampling design gives the smaller variance for the estimate of the number of employed individuals and does not have any effect to the estimates of the number of unemployed in comparison with the corresponding estimators which do not use the auxiliary data.

References

Aires, N. (1999). Algorithms to find exact inclusion probabilities for conditional poisson sampling and pareto πps sampling designs. *Methodology and Computing in Applied Probability* **1:4**, 457 – 469.

Bondesson, L., Traat, I. & Lundqvist, A. (2006). Pareto sampling versus conditional poisson and sampford sampling. *Scandinavian Journal Statitics* **33**, 699 – 720.

Horvitz, D. & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **78(381)**, 663 – 685.

Rosen, B. (1996). On sampling with probability proportional to size. *R&D Report 1996:1, Statistics Sweden* .

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

Sample survey of wages on professions and positions: experience of carrying out in the Republic of Belarus

Katsiaryna Chystsenka[1]

[1]National Bank of Republic of Belarus, e-mail: katsiaryna.chystsenka@gmail.com

Abstract

In this paper two-level sampling are considered. At the first stage are sampled enterprises, at the second stage are sampled workers. Weighting on two stages are considered. The basic sampling results are shown.

Keywords: Sample, wages on professions and positions, weighting, results of sample.

1 Introduction

Sample survey in sphere of labour statistics plays the important role for supervision of an economy condition and a population standard of living, survey of indicators of qualitative structure of wages. In Belarus the branches sampling of the enterprises have started to be carried out since 2006 in sphere of labour statistics:

The survey for the purpose of distribution of number of workers on the sizes of wages was carried out annually since 2007, since 2009 once in two years. However since November 2011 this survey is caring out continuous method.

The survey of structure of number of workers on sex, educational level, training and age in small enterprises was carried out in 2007 and 2008. Since 2009 the National statistical committee of the Republic of Belarus planned to carry out sample survey once in two years. However in 2009 committee substituted a sample method of survey on a continuous method.

One more sampling in sphere of labour statistics is sample survey of wages on professions and positions. The International Labour Organization recommended the realisation of this survey in Republic of Belarus. Since 2006 the National statistical committee of the Republic of Belarus carries out survey two times in five years.

Since 2012 the National statistical committee of the Republic of Belarus carries out quarterly sample survey of households for the study of employment problems.

Unfortunately now in Belarus sample survey is carried out only for the analysis of wages on professions and positions and for the study of employment problems in sphere of labour statistics.

Only the wages survey on professions and positions investigates the differentiation of wages. Results of this survey are necessary for international comparisons of purchasing capacity of wages on correlation of data about monthly wages of workers of surveyed professions with the prices of the basic food goods, for transition working out to new system of a payment with the account complexity, work intensity of workers and differentiation of wages on professions and positions.

2 Sampling design

Two-level sampling is used for survey of wages on professions and positions. At the first step territorial multivariate simple casual sampling of the enterprises is formed in the National statistical committee of the Republic of Belarus, at the second step workers are sampled by mechanical sampling algorithm in each enterprise.

2.1 The first stage

At the first stage objects of statistical survey are enterprises, their isolated divisions with number of workers more 16 persons of all patterns of ownership and small enterprises of not state pattern of ownership are an exception.

Sampling frame is a list of the enterprises representing the monthly state statistical report on labour (general aggregate of the enterprises). The enterprises with very small number of workers (till 4 workers) are excluded from the sampling frame for elimination potential non-responses.

The sampling frame is stratified into homogeneous groups of the enterprises by:

- regions of Belarus − 6 areas and Minsk-City and then
- economic activities − mining industry; manufacturing; production and distribution of electricity, gas and water; construction; transport and communication.

In each region of Belarus the small number of the enterprises is engaged mining industry therefore they were surveyed all.

Then sample of the enterprises is carried out in each region and economic activities contemporaneously on two indicators − wage fund and number of workers.

For survey can be used the combination of univariate sample (for large non-uniform general aggregate (over 400-500 units) and for very small size of a general aggregate (about 30-40 units)) and multivariate sample (for average non-uniform aggregate). The cluster analysis can be used for multivariate sample. For this survey was used the multivariate sample, but the cluster analysis was not used, because the stratified aggregates included a small number of enterprises.

The stratified aggregates were divided into uniform groups by the indicator «wage fund» for reduction of a sampling error. Sample is not carried out in groups with small number of the enterprises (till 6 units).

At the same time, the liquidation or temporary stoppage of activity of the enterprises often occurs in the process of sampling. Thus it is necessary to exclude the risks-enterprises from the general aggregate for reduction of non-responses. Sample updating is carried out taking into account several principles:

- risks-enterprise is replaced by other enterprise with approximately identical wage fund of the risks-enterprise; reweighing procedure is not executed;

- risks-enterprise is not replaced by other enterprise in two cases: 1) the wage fund of risks-enterprise is small and will slightly affect on the sampling error; 2) no enterprises to replace in corresponding group of the enterprises in a general aggregate; reweighing procedure is executed;

- if the number of the risks-enterprises makes more than 15 units in surveyed region and an economic activity, new sample of the enterprises is carried out in this region and an economic activity.

2.2 The second stage

At second stage the objects of statistical supervision are workers of sampled enterprises at the first stage. The sampling frame is a list of workers who worked full surveyed month (October).

Workers were ranged under personnel number of the worker. The name of professions and positions of workers are taken from the Nation-wide classifier of the Republic of Belarus «Professions of workers and positions of employees». It contains about 15 thousand professions and positions. Therefore the list of workers is stratified by categories: heads, experts, other employees, laborers.

The required number of sample workers is necessary to calculate. It number varied from 8 to 64 people and depends on the number of all surveyed workers. Sample is carried out in each layer (*i*) through the interval (*S*):

$$S = \frac{M_i}{m_i} \tag{1}$$

Here M_i – workers who worked full surveyed month (October),

m_i – the required number of workers for sample.

Number of initial worker is a random number (*RN*). It is chosen on drawing lots. All subsequent numbers are calculated as:

the second number:

$$RN + S \tag{2}$$

third:

$$(RN \cdot S) + S \tag{3}$$

fourth:

$$(RN \cdot 2 \cdot S) + S \tag{4}$$

last number:

$$RN + S\,(m_i - 1) \tag{5}$$

Economists of each sampled enterprises fill special state statistical report «The report on wages of workers on professions and positions for October». As a result of the previous survey report form has been changed. Additional column "Number of workers in this profession (position) who worked full surveyed month (October)" was added. In a previous survey the individual worker weight was calculated on categories of workers: heads, experts, other employees, laborers. But this method increased the sampling error. Additional column added for calculate of the individual worker weight on each profession (position).

The report consists of two parts. The first part contains information on the number and wages of all workers on categories. The second part contains the data stratified by professions for all surveyed workers: sex, education, experience, wage fund, number of worked hours and number of workers in this profession (position) who worked full surveyed month (October).

The report is filled from documents of the primary account and is sent to statistical offices.

3 Weighting and extrapolation

In the National statistical committee of the Republic of Belarus the obtained report data have been systematized, extrapolated and estimated at republican level on means of the special software developed for this survey.

Extrapolation is carried out with application of the aggregate weigh for a finite unit (worker):

$$x_{eij} = x_i \cdot k_{ai} \tag{6}$$

Here x_{eij} – the extrapolated value of x for i-th profession (position) of workers on j-th enterprise,

x_i – value of x for worker i-th profession (position),

k_{ai} – aggregate weight for i-th profession (position). Calculated as:

$$k_{ai} = k_e \cdot k_w \tag{7}$$

Here k_e – enterprise weight:

$$k_e = \frac{N_{ij}}{n_{ij}} \tag{8}$$

N_{ij} – number of the enterprises in a general aggregate for i-th economic activity of j-th region,

n_{ij} – number of the enterprises in a sample for i-th economic activity of j-th region,

k_w – individual worker weight for i-th profession (position). Calculated as:

$$k_w = \frac{T_i}{t_i} \tag{9}$$

T_i – total number of workers of i-th profession (position) who worked full surveyed month,

t_i – sampling number of workers of i-th profession (position).

4 Results

Results of sample survey of wages on professions and positions in 2011 have shown:

- as a whole on economic activities:

the sampling fraction varied from 21,4% to 41,7%,

the sampling error varied from 0,0% to 3,1% for indicator «wage fund», from 0,1% to 3,7% for indicator «number of workers»,

- as a whole on regions of Belarus:

the sampling fraction varied from 26,7% to 33,1%,

the sampling error varied from 0,2% to 1,2% for indicator «wage fund», from 0,1% to 2,0% for indicator «number of workers»,

- as a whole on the Republic of Belarus:

the sampling fraction amounted 29,5%,

the sampling errors amounted 0,2% for indicator «wage fund», 0,5% for indicator «number of workers».

By results of sample survey the bulletin is formed in National statistical committee of the Republic of Belarus. But it is used for office using and does not take places on an official site of statistical committee.

However results of survey in 2011 are not comparable with previous results due to the transition from a survey of industrial branches to the survey of some economic activities.

GÉSA, the survey control system in Hungary Frame, data collection, paradata and quality

Ildikó Györki[1]

[1]Hungarian Central Statistical Office, e-mail: ildiko.gyorki@ksh.hu

Abstract

Supporting the flow of survey design and data collection, the HCSO uses a standard metadata driven system, the so-called GÉSA system. This survey control system manages all economic and social statistical data collections of the office, observing the businesses and other institutions. The presentation introduces the central place of the system in the processing flow, the connection between registers, master frame and frames, the role of the system in the vertical and horizontal integration of different surveys. It highlights that the unified attributes and paradata make possible the standard monitoring, evaluation and quality assessment of the different surveys.

Keywords: survey, paradata, quality, frame

1 Introduction

In Hungary a general metadata driven system stands in the centre of the data collection systems from the middle of nineties. This so called GÉSA system manages all economic and social statistical data collections. The main principle of the system: the surveys, questionnaires are described in the metadata base, the survey frames have a common structure, during the data collection the same types of paradata are maintained with the same value sets (nomenclature), the functions are standardized, driven by metadata. This assures the unified monitoring and evaluation of the data collections and the effective development and operation of the system. The data collection, data editing and data processing functions are in close connection with the GÉSA system and survey frames. Because of the development of data collection instruments the importance of the different functions of GÉSA has changed during the almost twenty years, but the central role of the system and the unified approach of data collections remained steady.

2 The place of GÉSA in the processing flow

GÉSA (economic organisations and their data provision) is the survey control system for the observation of businesses and other institutions. It is the earliest metadata driven system in the HCSO which has been working since 1996. It is intended to give a general tool for supporting the tasks of the data collection phase of surveys, such as design, documentation, gathering of the questionnaires, and evaluation of data collections. The solution is built on standard procedures.

At present, GÉSA manages and controls 132 data collections which account for 98 percent of all data collections where the data suppliers are institutions. In the case of more than 80 surveys, the questionnaires can be filled in and sent to HCSO via internet with proactive support. Besides data collections, 50 administrative sources belong to GÉSA for the sake of unified processing. The system maintains a population with almost 2 million units, more than 380 000 data suppliers and 2-3 million pieces of questionnaires in a given year.

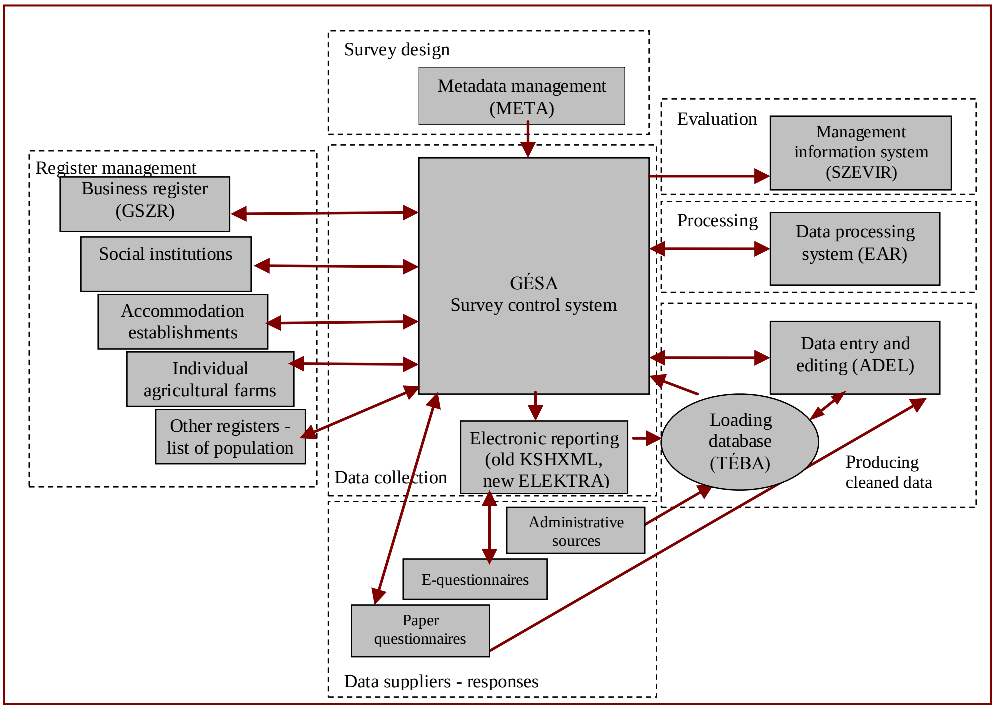
The GÉSA is built on the Business Register, its satellite and supplementary registers. In Hungary the Business Register covers the legal units with tax number (more than 1.7 million existing units), their local units and kind of activity units for the corporations over 50 employees. There are different other specialized satellite registers that are connected to the Business Register via the tax number like retail trade, social institution register, accommodation register, healthcare, research and development register, etc. One part of the satellite registers manages local kind of activity units. There are supplementary registers to the Business Register that contains not only the units with tax number but units without tax number like non-profit register with non-profit institutions and farm register with individual farms. These units without tax number are additional units to describe the whole population of economy. The GÉSA manages the surveys with population built on these registers. The interview type surveys built on the address register (population surveys) don't belong to the GÉSA, they are managed by the other survey control system (named LAKOS).

During the survey design the structural metadata of surveys have to be described every year: survey identification data, way and instrument of data collection, exact scope and forming rule of population, data supplier and statistical unit of the survey, sampling plan, way of mailing the questionnaires, detailed description of the questionnaires for personalization, deadlines of the phases of the data collection, etc.

The GÉSA functions are built on these metadata. The functions are performed for each reference period. The GÉSA supports the data collections via post, e-mail and web, the interview type data collection (for agriculture), and those secondary sources where the population is known in advance. The main principle of the functions is to give proactive support for the data suppliers to fulfil their duties. For the statistician colleagues whose task is the data collection the GÉSA gives support to deal with the subpopulations belonging to their responsibilities. The most important functions are the following:

- Forming the master frame from the snapshots of the registers (see the next chapter)
- Assigning the survey frames of the different surveys based on the master frame
- Selecting the sample of the representative data collections
- Selecting the questionnaires for mailing, their personalization, creating control information for printing, making a calendar for the data suppliers with their own response deadlines.
- Automatic and manual reminding and urging the data suppliers for response according to the urging plan by e-mail, fax and letter.
- Registering the responses, the way of responses, the non-responses and the negative answers and their cause.
- Maintaining and feed backing the changes and errors in the contact and other register information to the registers.
- Monitoring the flow of the data collection, to evaluate the result
- Computing quality indicators and response burden

Figure 1: The environment of the GÉSA system – relation among the statistical processing



Survey design
Metadata management (META)
Evaluation
Management information system (SZEVIR)
Register management
Business register (GSZR)
Social institutions
Accommodation establishments
Individual agricultural farms
Other registers - list of population
GÉSA Survey control system
Processing
Data processing system (EAR)
Data entry and editing (ADEL)
Loading database (TÉBA)
Producing cleaned data
Electronic reporting (old KSHXML, new ELEKTRA)
Data collection
Administrative sources
E-questionnaires
Paper questionnaires
Data suppliers - responses

GÉSA is in close connection with the different data collection techniques. For electronic reporting (primarily reporting on web now with the KSHXML, from the next year with the ELEKTRA system) the connection is direct, the GÉSA gives the duties of data suppliers, the frame information, statistical units, collecting units, personalisation and contact information for the respondents and for their e-questionnaires.

The questionnaires provided in an electronic way (web, e-mail, secondary source) are loaded into the data base. Their validation control and registration in the GÉSA is also unified and automatic (TÉBA system).

The paper questionnaires are entered to the data base by a frame system, ADEL. Its task is not only the data entry but it manages the data loaded from electronic source as well. The aim of this phase is to produce cleaned data. For the control, validation and editing of data the ADEL is in direct connection with the GÉSA.

The processing phase for imputation, estimation, aggregation and other analysis functions also use the survey frame information managed by the GÉSA.

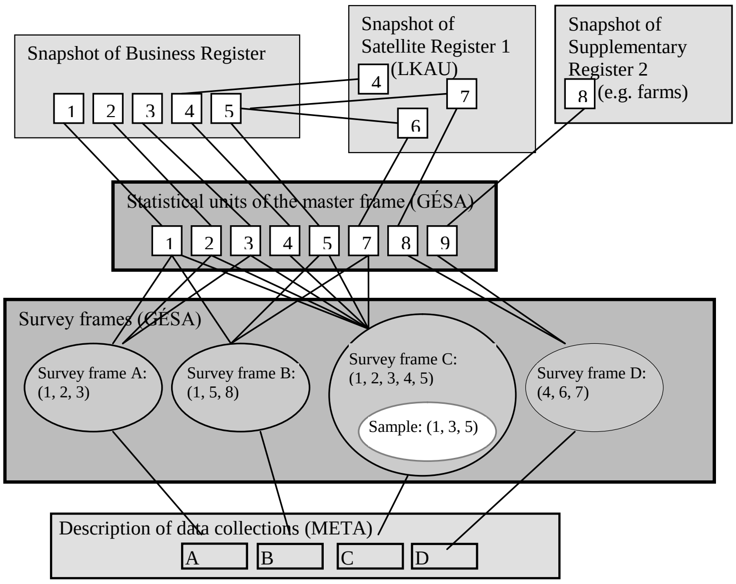
In order to evaluate the surveys, to analyse the quality, to feedback the information to the survey design phase the GÉSA provides unified information for all surveys managed by it. It gives statistics and indicators about the data collections automatically and as frequently as it is demanded.

3 Survey frames and data integration

The frames of economic and social statistal surveys are created by the GÉSA system. The assigment of the data suppliers and statistical units is built on the *master frame* and metadata. The master frame is created according to the reference period of the surveys in every month and at the end of the year from the register snapshots referring to the given time. At forming the master frame the most important element is the Business Register, but the master frame contains not only the statistical units of the Business Register but the statistical units of its satellit and supplementary registers as well. The united statistical unites are labelled with unified identification number and statistical unit type.

During the survey design phase a precise description is made in the meta database about the scope and units of the frame population (data suppliers, statistical units) and the connection with other surveys. The population is united from one or more subpopulations. The subpopulation can be selected by an algorithm on the attributes of the master frame unit. Beside that, other sources can be used to the assignment as well, like the result (data), experience (paradata) of the same or another survey from the previous period, or external sources.

Figure 2: Creation of the master frame and the survey frames


Snapshot of Business Register
1 2 3 4 5
Snapshot of Satellite Register 1 (LKAU)
4
7
6
Snapshot of Supplementary Register 2 (e.g. farms)
8
Statistical units of the master frame (GÉSA)
1 2 3 4 5 7 8 9
Survey frames (GÉSA)
Survey frame A: (1, 2, 3)
Survey frame B: (1, 5, 8)
Survey frame C: (1, 2, 3, 4, 5)
Sample: (1, 3, 5)
Survey frame D: (4, 6, 7)
Description of data collections (META)
A B C D

It is an important task of the survey design to prepare the integration of the survey data with the proper definition of the subpopulation of the surveys. We differentiate horizontal and vertical integration.

The first means linking statistical measures from different sources for a given population. For example, sales data of a retail trade data collection can be linked to the data of a labour survey. For the horizontal integration has to define common population (subsets of the population) for the surveys. These subsets are described and identified in the metadata base. The assignment of the population provides the same units for the subsets of the topics that we are planning to integrate. It is practical to select a common sample for the common subsets of the surveys because it helps the comparison not only for the estimated but the sample data as well.

In the case of vertical integration, we make a union of a given statistical measure for different, separately collected subsets of a particular population. For example, unifying data on the land usage of agricultural organizations, collected by self-enumeration with those of individual agricultural farms, collected by interviews creates data for the whole national economy. The base for the vertical integration of data is the definition of disjunctive populations of the surveys.

The selection of the survey frames from the master frame is automatic, built on the description of the population in the metadata base and the sources described. The assignment takes into account the relation of the surveys and the standard subsets of populations described above.

The survey frames and samples are stored and managed in one database table with common attributes for a reference period, so it gives an easy opportunity to analyse the survey frames, and the response burden. There is an application within the GÉSA system that makes data mining possible in order to analyse the different strata and units of the all populations together according to the response burden.

4 Unified paradata, monitoring and quality indicators

Information about the statistical units and data suppliers, the fault in their main attributes coming from the registers, the coverage error of the population, the important steps and the success of the data collection are collected and registered during the data collection on statistical unit level for all surveys and for all way of data collection methods. Where it is possible the information on the data collection phases is standardized, they are described by unified code lists (nomenclatures). In other cases textual information can be added to the data suppliers and statistical units about their readiness for response, and the changes in the unit that can be useful for the validation of the responses.

The most important attributes, paradata to monitor and evaluate the data collection phase of the surveys are the excepted and realized way of response (paper, e-mail, web), the steps of urging, the type of response (response with data and response without data), the reason of non-response or negative answer. The non-responses can be ranged into three groups:

- the first group deals with the frame problems: status of the statistical unit, data supplier (dead, under liquidation, etc.), classification problems, accessibility problems
- the second group deals with activity problems of the statistical unit, that has not the observed activity (now, ever, temporarily…)
- the third group deals with the respondents (it denies, is overdue, no successful contact, unknown cause)

During the data collection each questionnaire is characterized with these codes, if it is missing or it is sent without data. These codes are used at imputation and estimation to state who belongs to the survey population and who is not.

The information gathered in the data collection phase is used:

- for monitoring the progress of the data collection phase, to organise and control the work, to time the urging and other steps.

- for feed backing information about the register errors to the suit register.

- For determining the over coverage (unit doesn’t belong to the population) and under coverage (unit must belong to the population, but formerly there was no information on it), and using in the next round of the data collection phase at the assignment of the survey frame.

- for evaluating the result and quality of the data collection.

There are statistics about the data collections automatically produced every day and every month. These inform about the rate of the different data collection instruments (paper, web, e-mail), the way of data entry (loaded data, manual data entry), the rate of the non-response. The detailed reports about the reasons of non-response, urging types, imputations, etc. are queried by the users of the GÉSA application.

The accuracy indicators of the data collection quality about the coverage, frame errors, the non-response, unit level imputation can be computed automatically.

References

Györki, I. (2012). GÉSA: The Tool for Survey Control, Quality Assessment and Data Integration. *Hungarian Statistical Review, Special number 15,* **48-78.**

Analysis of repeated business surveys in Ukraine

Oksana Honchar[1] and Tetiana Ianevych[2]

[1]Scientific & Technical Complex for Statistical Research, National Academy of Statistics, Accounting & Audit, e-mail: ohonchar@list.ru
[2]Taras Shevchenko National University of Kyiv, e-mail: yakovenkot@gmail.com

Abstract

In the paper methodology of repeated surveys is consider on example Ukrainian investment survey. Short characteristic and sample design of quarter investment survey are presented. The allocation depending on the fraction α of the sample size is considered. Horvitz-Thompson and regression estimator using information from previous survey are compared using simulation study.

Keywords: Investment survey, repeated surveys, Horvitz-Thompson estimator, regression estimator, bias, mean square error.

1 Introduction

Most surveys in short-term business statistics are repeated with some periodicity. In most cases they are conducted monthly or quarterly but sometimes also have weekly or two-month periodicity. For such surveys special methodology should be used.

As an example we consider the Ukrainian investment survey, which plays an important role as a source of data regarding capital investment, along with its components, and provides the data for constructing the capital investment index, which is one from the main indicators of economical statistics. The capital investment survey is conducted with annual and quarter periodicity. The main aim of annual survey of capital investment is estimation of the structure whereas quarter survey has estimation of changes as an objective.

All enterprises with capital investment in a reporting year should be observed. In the quarter survey only enterprises that have significant size of capital investment are observed. Actually only enterprises which took part in other quarter surveys (mostly in structural business survey) can be observed quarterly.

The first problem in the quarter survey is in the future frame for this survey will be formed on the central level (in SSSU). This population will be formed from the business register. In 2010 the frame included 689 042 enterprises. For the quarter survey some threshold is used to take part in the survey. Until now it was defined by capital investment itself since before to obtain questionnaire two questions were asked to a potential respondent: 1) if enterprise had capital investment in this quarter and if yes 2) what was size of capital investment. If size exceeded some threshold (different by activities) an enterprise was observed. However forming of population on the central level does not presume knowledge of capital investment size so using it for threshold definition is impossible.

The second problem concerns with quality of the quarter data from the investment survey. Quarter capital investment was not estimated for enterprises that were not observed. So we have a problem of underestimating capital investment size in quarter surveys. It can be no problem for capital investment index calculation because size of underestimation is proportional to proper size of capital investment by domains. But it becomes problem using absolute values of capital investment.

2 Sampling design

The main characteristics of Ukrainian investment survey are presented in Table 1.

Table 1: Characteristics of Ukrainian investment survey

Kind of economic activity	• All kinds
Geographical coverage	• All regions in Ukraine (27)
Unit	• Enterprise
Periodicity	• Quarter; • Annual
Main variable	• Total capital investment
Kind of survey	• quarterly: – now: cut-off with census; – plan: cut-off with sample: census for large and middle enterprises; sample for small enterprises; • annually: census
Levels of publication	• National level; • Regions; • Kinds of economic activities (2digit by NACE); • Institutional sector of economy; • Organization and legal form of management

Sample frame for quarter investment survey is formed from the business statistical register. Then some data is added from the previous year's annual structural business survey and investment survey for sampling design construction. Sample frame consists of three subpopulations: financial sector of economy, non-financial sector and new enterprises. For sample survey small enterprises are separated in terms of financial and nonfinancial sectors with all new enterprises covered regardless of size.

For small enterprises stratification by activity (2-digit of NACE) and numbers of employees (0, 1-2, 3-5, 6-9, 10-19, and 20-50) is used. Then outliers are detected. Enterprise is atypical (outlier) if the condition $\left|y_{hi} - y_h\right| < 3 \cdot \sigma_h$ does not hold. Here y_{hi} – capital investment for unit i in stratum h, y_h – mean of capital investment in stratum h, σ_h – standard deviation of capital investment in stratum h.

Total sample size is about 100000 units (Table 2).

Table 2: Sample size in Ukrainian investment survey in 2010

Subpopulation	Size	%
Large & middle enterprises	35 139	100
Outliers and small strata	2 088	100
Small enterprises observed with p <1	59 292	9.5
New enterprises	3 305	10
Total	99 824	14.5

Sample with such design gives quite good results for total investment but the problem of small domain estimation still exists. For estimation Horvitz-Thompson and regression estimates are calculated. As auxiliary information, the number of employees in current period is used.

3 Using information from previous survey

As long as this is repeated survey, for improvement of estimates accuracy, as auxiliary information, we decided to use capital investment from previous year's survey.

Resources of investigation are:

1. Quarter data on capital investment in 2010;

2. Annual data in 2009, 2010;

3. Statistical register data in 2010.

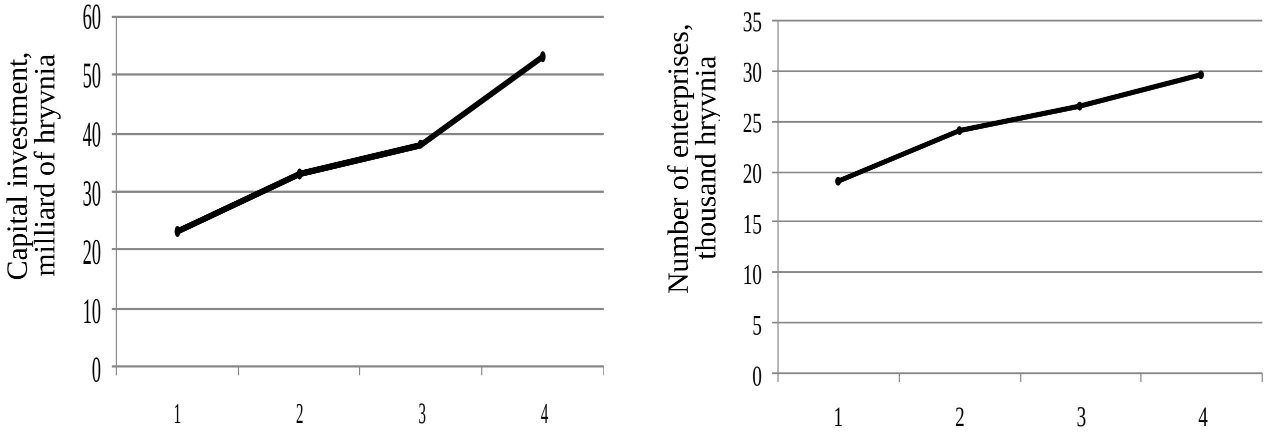
Table 3: Distribution of enterprises by capital investment size by quarters in 2010

Period	Capital investment, thousand hryvnia	Rate of quarter capital investment in annual investment, %	Number of enterprises
quarter1	23 220 146	13,7	19 017
quarter 2	33 432 883	19,7	23 658
quarter 3	38 397 367	22,7	26 416
quarter 4	52 946 719	31,2	29 590
Sum of quarters	147 997 115	**87,3**	×
Annual survey	169 434 303	100,0	71 352

Table 4: Distribution of enterprises by quarters in 2010

Criterion	Number of enterprises	Rate of enterprises, %
Enterprises which were observed:		
in one quarter	14 317	19,9
in two quarters	8 563	11,9
in three quarters	8 138	11,3
in four quarters	10 706	14,9
only in annual survey	30 354	42,1
Number of enterprises reported in annual survey	72 078	100,0

Capital investment and number of enterprises by quarters in 2010



First we compare the behavior of the H-T and regression estimator on the population of all enterprises. It was divided into two parts: those that had capital investment in 2009 and those that not.

These subpopulations are essentially different. So, in 2009 93% enterprises had not capital investments and it can be concluded from the data we have, that it is most likely that the enterprises that had not investment during one year would not have it during the next. The subpopulation of such enterprises is very homogeneous (it consists mostly of zeros) therefore, the use of the H-T estimator is appropriate there.

The subpopulation of enterprises that have positive capital investment in 2009 is not so homogeneous. So, for the beginning we compared the behavior of the H-T and regression estimator on this subpopulation within simple random sampling. Besides this, we investigated what is the best allocation of the sample between this subpopulation. Let us consider 4 cases of the allocation depending on the fraction α of the sample size n.

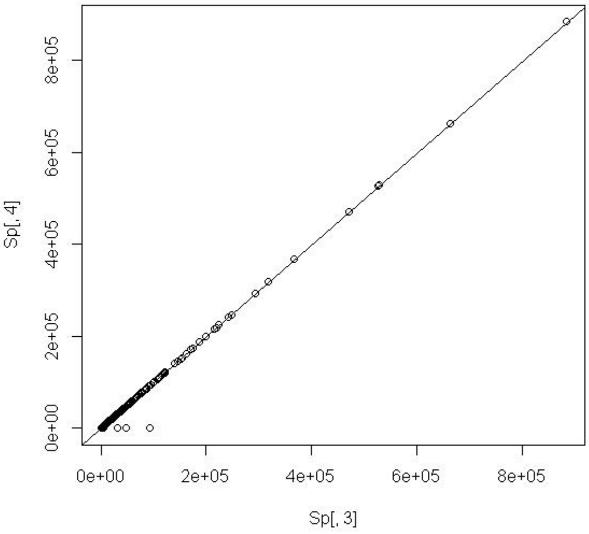


Table 5: Sample allocations

Sample sizes	α=0.5	α=0.6	α=0.7	$\alpha\approx$0.77
n_p	29646	35575	41504	45598
n_0	29646	23717	17788	13694

In the Table 5 the value n_p is the sample size for the subpopulation of enterprises that had capital investment in 2009, and the value n_0 is the sample size for the subpopulation with zero investment. The last case ($\alpha\approx$0.77) corresponds to the situation when the enterprises with nonzero investments in 2009 is observed with probability one.

The parameter t that we are interesting in is the total capital investment for small enterprises in 2010. The estimates of this parameter has two terms $\hat{t} = \hat{t}_p + \hat{t}_0$, where $\hat{t}_0$ is the estimate for the total investment for subpopulation of enterprises with zero investment in 2009. For $\hat{t}_0$ we only considered the H-T estimator. $\hat{t}_p$ is the estimate for subpopulation with nonzero investment and for this term we considered two alternatives – H-T and regression estimators. As an auxiliary variable for the regression estimator we used capital investment for the enterprise in the previous 2009 year. For annual surveys this data is strongly correlated, so we may wait for the essential improvement using the regression estimator. The picture shows the almost ideal linear dependence between 2009 and 2010 data.

Estimates for quarter survey will be presented at the workshop.

4 Simulation study

For comparison accuracy of estimators Monte-Carlo method with K = 10000 simulations was used. As accuracy indicators absolute relative bias $ARB = \left| \frac{1}{K} \sum_{i=1}^{K} \hat{y}_d(s_i) - Y_d \right| / Y_d$ and relative root mean squared error $RRMSE = \sqrt{\frac{1}{K} \sum_{i=1}^{K} (\hat{y}_d(s_i) - Y_d)^2} / Y_d$ are calculated. Simulation results are presented in Table 6.

Table 6: Comparison of the estimators

Estimator	Accuracy indicator	α=0.5	α=0.6	α=0.7	$\alpha \approx$0.77
$HT_p + HT_0$	ARB,%	0.09	0.03	0.02	
	RRMSE,%	4.03	2.97	1.87	
					0.007
$GREG_p + HT_0$	ARB,%				0.69
	RRMSE,%	0.01	0.009	0.01	
		0.52	0.53	0.62	

Remark. In last case ($\alpha \approx$0.77) the bias and the variation are caused only by estimation of enterprises with zero capital investment in 2009.

Estimation by domains is also very important for capital investment survey. Results of studying by activities and regions will be presented at the workshop.

References

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

Initial Wave Nonresponse and Panel Attrition in the Finnish Subsample of EU-SILC

Tara Junes[1]

[1]Statistics Finland, University of Helsinki, e-mail: tara.junes@stat.fi

Abstract

The objective of this paper is to study the effects of initial wave unit nonresponse and panel attrition on the quintile distribution of disposable household equivalised income. Analyses are performed for one rotation group selected from Finnish EU-SILC. In addition to empirical analysis the changes between quintiles states are modelled with Markov chains.

Keywords: Unit nonresponse, Attrition, EU-SILC, Markov chains

1 Introduction

This paper is a summary of my Master's Thesis (Junes 2012). The purposes of the study was to investigate unit nonresponse in the Finnish subsample of EU Statistics of Income and Living Conditions (EU-SILC) which is a panel study with a four-year rotational sampling design.

The common perception is that estimation results of a panel study become more biased with increasing amount of attrition. However, it has been shown that a nonresponse bias at the beginning of the panel can fade away in subsequent panel waves without any correction. The fade-away phenomenon occurred in the analyses of certain income variables, from which the most interesting is disposable household equivalised income. (Rendtel et al. 2004 and Gerks 2004).

The fade-away theory was suggested by analysing the Finnish subsample of European Community Household Panel (ECHP). The objective of my Thesis was to investigate the fade-away hypothesis for a different dataset and to show that the existence of the hypothesis is not so straightforward. The dataset being analysed consists of information collected for one rotation group from Finnish EU-SILC. The selected rotation group can be seen as a non-rotational panel with duration of four years the initial wave being the first analyse year.

The main attention is given to computation of the income quintiles and to the empirical and theoretical modelling of transitions between computed quintiles. The research is performed for three different groups of sampling units:

1 All sampling units that were intended to participate the panel at the initial wave i.e. in 2005 (FULL-sample).
2 All respondents of the initial wave (RESP-sample).
3 All observed panel members in subsequent waves (OBS-sample).

The effects of initial wave nonresponse are displayed by comparing the income distribution of FULL-sample with the corresponding results from the RESP-sample. If there is differences between the RESP-sample and the OBS-sample, it is a possible sign of attrition bias.

The main analysis variable is disposable household equivalised income which is the total gross household income minus current transfers paid adjusted by the household composition. The adjustment is done with the OECD-modified equivalence scale assigning a value of 1 to the first household member, of 0.5

to each additional adult and of 0.3 to each child (Atkinson *et al.* 2002). The total disposable household income of a dwelling unit is divided by the sum of the scaling values after which the quotient is assigned to each individual in the household. In the previous analyses fade-away effect was seen in the distributions of both equivalised and non-equivalised household income (Gerks 2004).

2 Data

The Finnish EU-SILC sample 2006 was drawn from the Population Information System (PIS) maintained by the Population Register Centre of Finland. The number of persons belonging to the sample for the selected rotation group is 2 500. The rotating structure of the panel is displayed in Table 1. The income information for the sampling units was collected from the registers of Statistics Finland.

Table 1: Rotation structure of EU-SILC

	Measurement year								
Rotation group	1	2	3	4	5	6	7	8	...
R1	X	X	X	X					
R2		X	X	X	X				
R3			X	X	X	X			
⋮									

The research data were created by linking the household dwelling unit to the sampling unit and subsequently adding the household income information of the dwelling unit. This procedure was iterated for all waves of the panel so that the changing household composition was taken into account. All analyses use register information for respondents and nonrespondents alike.

The dataset of 2 500 sampling units includes also persons not belonging to the study population and hence the total sample size was slightly reduced. However, even after the exclusion of the over-coverage there were still persons creating difficulties in the analyses: for some households there were no household income information available in the household registers of Statistics Finland. For attaining the comparability between the former analysis performed for ECHP and current analysis performed for EU-SILC, persons having missing values in the income variables were excluded from the analyses of EU-SILC dataset also (Gerks 2004).

The total amount of 2 353 households had their income available in the registers in all analysed income reference years and thus it is also the starting point of the analyses of this paper. From now on this subset of the original sample is referred as the FULL-sample. The total amount of respondents to all four waves is 1 448 and the number of persons responded at the initial wave is 1 769. Thus with the FULL-sample of 2 353 respondents the amount of initial wave nonresponse is 584 persons being round 25 percent. From now on the respondents to all four waves are referred as the OBS-sample (obs as observed) and respondents at the initial wave are referred as the RESP-sample (resp as respondents).

2.1 Quintiles of disposable household equivalised income

The main attention of this paper is given to analysis of transitions between income quintiles of the disposable household equivalised income. The transition analysis bases on the FULL-sample and hence it is enough to analyse the income quintiles of the FULL-sample only. As it was mentioned previously the income reference period is always the year preceding the survey year, that is income distributions of EU-SILC 2006 base on income data from reference year 2005 and so on. Income quintiles for the disposable household equivalised income are displayed in Table 2.

Table 2: The disposable household equivalised income, FULL-sample

Year	N	The 20 th	The 40 th	The 60 th	The 80 the
2005	2 353	12 720	17 147	21 302	28 119
2006	2 353	13 016	17 708	22 492	29 550
2007	2 353	13 668	18 907	23 854	31 392
2008	2 353	14 384	19 524	24 777	32 130

If the respondents were allocated to quintiles according to the percentile points of the analyse year, 20 % of the respondents would be in every quintile in every analyse year. This is of course the idea of quintile computation but it does not help in the analysis of transitions. Thus the percentile points of the income reference year 2005 are selected as fixed percentile points for all subsequent panel waves.

An adjustment to the percentile points is still required because of the inflation, for instance. The idea is to adjust the percentile points by the ratio of the median of the analyse year and the median of the base year 2005. Adjusting by the median ratios is a method Gerks (2004) applied in the analyses of European Community Household Panel in order to prevent the clustering of respondents into a one quintile. The medians and median ratios are displayed Table 3. The income bracket for quintile number one in 2008 is $1.14 \times 14384 = 16397.76$, for instance.

Table 3: The medians and median ratios for the FULL-sample

Year	N	Median	Median 2005	Median ratio
2005	2 353	19 322	19 322	1.00
2006	2 353	19 956	19 322	1.03
2007	2 353	21 326	19 322	1.10
2008	2 353	22 054	19 322	1.14

3 Markov chains

3.1 Theory

In general, let $\{X_t, t = 0, 1, 2, \ldots\}$ be a discrete time stochastic process with finite state space $E = \{0, 1, \ldots, N\}$. If the conditional probabilities at time $t+1$ satisfy

$$P(X_{t+1} = j | X_0 = i_0, X_1 = i_1, \ldots, X_t = i) = P(X_{t+1} = j | X_t = i), \tag{1}$$

the stochastic process is called a Markov chain. The property 1 is also known as the Markov property or the memoryless property. (Brémaud 1999)

The conditional probabilities are called transition probabilities and they are collected into the matrix $\mathbf{P} = \{p_{ij}\}_{i,j,\in E}$, where

$$P(X_{t+1} = j | X_t = i) = p_{ij}. \tag{2}$$

If the probabilities defined in equation 2 are independent of the time point t, the Markov chain is said to be time homogeneous. The transition probabilities sum to one, i.e. $\sum_{j=1}^{k} p_{ij} = 1$, where $p_{ij} \geq 0$ for all $i, j \in 1, \ldots, k$. (Brémaud 1999)

The random variable X_t at time point $t = 0$ is called the initial state of the Markov chain with initial probability distribution ν given by

$$P(X_0 = i) = \nu(i), \tag{3}$$

where $i \in E$ and $\sum_{i=1}^{k} \nu(i) = 1$. The initial distribution tells us how the Markov chain starts. The initial distribution and the transition matrix determine the distribution of the discrete-time homogeneous Markov chain. (Brémaud 1999)

The state of the Markov chains after t-steps is computed by using the t-step transition probabilities defined by

$$p_{ij}^{(t)} = P(X_{s+t} = j | X_s = i), \tag{4}$$

where $s \geq 0$ is the selected time point and t is the selected time interval or time step. Because of the homogeneity assumption of the chain, probabilities in equation (4) are independent of the value of s. (Häggström 2002)

The t-step transition probabilities are computed from the first step transition probabilities and are given by

$$p_j^{(t)} = \sum_{i \in E} \nu_i p_{ij}^{(t)}, \tag{5}$$

where $p_j^{(t)} = P(X_t = j)$ is the probability of being at state j after t time steps and $\nu_i = P(X_0 = i)$ is the initial probability of state i. The initial distribution, the distribution after t time steps and the transition probabilities are possible to present in matrix form. Hence equation (5) becomes

$$\mathbf{p}^{(t)} = \boldsymbol{\nu} \mathbf{P}^t, \tag{6}$$

where $\mathbf{p}^{(t)} = (p_0^{(t)}, p_1^{(t)}, \ldots, p_N^{(t)})$ is the t-step distribution, $\boldsymbol{\nu} = (\nu_0, \nu_1, \ldots, \nu_N)$ is the initial distribution and $\mathbf{P}$ is the matrix containing first step transition probabilities. (Brémaud 1999)

3.2 Computation of transition probabilities

Markov chain modelling is used for studying transitions between adjusted income quintiles. A person (representing his/her household) belonging to the FULL-sample has five possible income quintile states in every analyse year. Thus the Markov chain in question is a process $\{X_t\}$, where $t = 0, 1, 2, 3$ with a state space consisting of the quintiles $E = \{1, 2, 3, 4, 5\}$. Here quintile one is assigned to the lowest income class and respectively quintile five is assigned to the highest income class. If it is also supposed that the time has no effect on the first step transition probabilities, the process in question is a discrete time and a finite discrete state space homogeneous Markov chain.

Because the Markov chain in question satisfies the conditions of equation (1) the transition probabilities are computed by applying equation (2). The actual calculation is done with the ℓEM-software in which the estimation bases on maximum likelihood theory and on the EM-algorithm (Vermunt 1997). The estimated transition probabilities are displayed in Table 4. The transition probabilities are supposed to be the same for respondents and nonrespondents.

As expected transitions into the current state are more probable than into any other states. This phenomenon is perhaps greater in the top and bottom quintiles, where the probability at staying in the current state is over 75 %. This high probability at staying in the current state may have something to do with the fact that a household belonging to the lowest income quintile and having a decreasing amount of income every year cannot change its quintile position. The same applies to the uppermost quintile class when the word decreasing is replaced by increasing.

Table 4: The estimated transition probabilities

	End				
Start	1	2	3	4	5
1	0.7586 (0.0112)	0.1664 (0.0098)	0.0454 (0.0055)	0.0213 (0.0038)	0.0083 (0.0024)
2	0.1601 (0.0099)	0.5779 (0.0133)	0.1863 (0.0105)	0.0575 (0.0063)	0.0182 (0.0036)
3	0.0450 (0.0057)	0.1709 (0.0103)	0.5255 (0.0137)	0.2181 (0.0113)	0.0405 (0.0054)
4	0.0258 (0.0042)	0.0580 (0.0062)	0.1599 (0.0097)	0.6013 (0.0129)	0.1550 (0.0096)
5	0.0300 (0.0045)	0.0137 (0.0030)	0.0348 (0.0048)	0.1331 (0.0089)	0.7884 (0.0107)

4 Results

The results of the empirical and Markov chain analyses are collected into Table 5. The transition probabilities of the FULL-sample displayed in Table 4 are used for attaining the Markov chains results also for the RESP-sample. It seems that there is only small amount of initial wave nonresponse bias present in the analysed rotation group of Finnish EU-SILC. Because of lacking nonresponse at the initial wave there is no fade-away phenomenon present in the dataset. It is clear that attrition biases the results making the distribution into quintiles more skewed during the lifetime of the panel.

Table 5: Empirical and theoretical analysis results

	Year 2005		**Year 2008**				
	Full	**Resp**	**Full**		**Resp**		**Obs**
Sample size	2 353	1 769	2 353		1 769		1 448
Distr. on states	**Emp**	**Emp**	**Markov**	**Emp**	**Markov**	**Emp**	**Emp**
$p(1)$	20.0	19.3	20.8	20.4	20.5	20.5	18.9
$p(2)$	20.0	20.1	19.4	19.8	19.3	19.3	18.7
$p(3)$	20.0	20.0	18.4	18.7	18.4	18.2	18.1
$p(4)$	20.0	20.5	20.9	21.1	21.0	21.7	22.2
$p(5)$	20.0	20.1	20.6	20.1	20.7	20.4	22.1

From the Markov chain modelling point of view comparisons made between the simulated and empirical distributions in Table 5 are promising. The simulated distributions for the FULL- and RESP-samples are close to their empirical distributions and hence the usage of the Markov chain modelling is justified.

The computed transition probabilities could be used for simulating the future states of the chain to see what happens to the initial wave nonresponse bias with longer duration of the panel than four years. But because the selected sample contains no real initial wave nonresponse bias it is clear that simulation will not provide any support to the fade-away hypothesis. Hence the simulation results are not displayed here and the reader is advised to consult Junes (2012, p. 57) for the results.

References

Atkinson, T., Cantillon, B., Marlier, E. & Nolan, B. (2002). *Social Indicators: The EU and Social Inclusion.* Oxford University Press.

Brémaud, P. (1999). *Markov Chains: Gibbs Field, Monte Carlo Simulation, and Queues.* New York: Springer.

Gerks, H. (2004). *Zur Stabilität von Nonresponse-Effekten in Panelerhebungen.* Bachelor's thesis, Freie Universität Berlin.

Häggström, O. (2002). *Finite Markov Chains and Algorithmic Applications.* Cambridge University Press.

Junes, T. (2012). *Initial Wave Nonrespone and Panel Attrition in the Finnish Subsample of EU-SILC.* Master's thesis, University of Helsinki.

Rendtel, U., Behr, A., Bellgardt, E., Neukirch, T., Pyy-Martikainen, M., Sisto, J., Lehtonen, R., Harms, T., Basic, E. & Marek, I. (2004). Report on Panel Effects. Results of Work Package 6 of the CHINTEX-Project, CHINTEX. Retrieved 11.1.2012 from `http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Wissenschaftsforum/Chintex/ResearchResults/Downloads/WorkingPaper22,templateId=renderPrint.psml`.

Vermunt, J. K. (1997). ℓem: A General Program for the Analysis of Categorical Data. The Netherland: Tilburg University. Retrieved 3.5.2012 from `http://spitswww.uvt.nl/web/fsw/mto/lem/manual.pdf`.

Survey Organisation Issues for Micro and Small Size Enterprise Managers

Biruta Sloka[1] and Ināra Kantāne[2]

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

Abstract

Survey organisation issues for micro and small enterprise managers are on great importance for research related to micro and small enterprise management problems. Currently many of surveys are organised by help of internet, but, unfortunately not all micro and small size enterprises have access to internet in Latvia. For evaluation on different issues is used scale 1 – 10 to get more detailed information and to use descriptive statistics and multivariate statistics analysis for data processing.

Keywords: SME, survey organisation, webpage, response rate, evaluation scale

1. Introduction and theoretical background

Academic and practical research more and more pay attention to quality of survey organisation, realisation and survey data processing. Thousands of scientific publications are devoted to those issues. Researcher groups have evaluated how often, on what extent and how deep such research has been done. Questionnaire design, survey realisation, response rate, tools for organisation of responses collection, issues of representativity ensurance of the sample, and many other issues are on agenda for researchers. Current paper has examined some theoretical findings as well as practical survey realisation for micro and small enterprise managers in Latvia in 2010 – 2011. Research methods applied: academic literature studies and updates, evaluation of real survey realisation: comparisons for population and sample data, discussions on attitude measurement scale use.

Ardilly, P. & Tillé, Y. (2006) have developed sampling methods and realisation materials, Janes (1999) have devoted attention to survey construction, Lavallée, P. & Rivest, L.-P. (2012) have devoted attention to capture–recapture sampling and indirect sampling, Särndal, C., Swensson, B. & Wretman, J. (2003) have paid attention to model assisted survey sampling. Issues on questionnaire design and distribution have been on research agenda for many researchers, like Kelly, (2000), Black, *et al.* (2005). Many researchers and practioners have made discussions on most convenience and effective evaluation scales used for research: Chang, *et al. (*2001) has developed important arguments for 5 point scale use, Garrat, *et al.* 2011) have made extensive discussion what scale: with 5 points or 10 points are more effective, Verdegem, *et al.* (2009) has used 11 point scale for survey realisation. Melnik, *et al. (*2012) have indicated that only every third manager do not refuse respond for surveys. His team research has found that during the last decade the response rates for surveys have even declined. Froflich, *et al.* (2002) have studied how to increase response rates, Dennis (2003) has studied by realisation of experiment specific issues on increase of response rates for small and medium size enterprises, to specific issues on survey realisation in SME has been devoted Knight's (2001), Smith's, *et al.* (2007), Redoli, *et al. (2008)* research. Sivo, *et al. (*2006) have examined – how low you should go, minding the response rates and representation issues, Tsatsow, *(*2011), Ilieva, *et al.* (2002), as well as Dijk, *et al.* (2009) have examined that internet is making huge changes in survey organisation and realisation, such issues were on research agenda also in the last decade represented by Dillman's, (2000) research. Several studies have been devoted to different tools for survey data analysis, like Melnik, *et al.*(2012).

2. Survey organisation and main empirical research results

The population of the survey were micro and small medium enterprises in Latvia. Survey was conducted from December 2010 till August 2011 when started economic growth after financial crisis. Respondents were selected by systematic sample (to be able to use different multivariate analysis methods for data processing), it was approached every third company from Latvia Investment and Development Agency and LAD who have contracts in period 2007-2013 and every tenth company from ZL Hotline data basis. It was made sure to avoid inclusion of the respective company from different data bases. Before the survey it was made 8 pilot interviews to test the questionnaire. For survey mainly was used internet survey, telephone survey or interviews in case of unavailability of the internet access. At the beginning it was given a phone call to micro and small company managers to invite to participate in the survey and fill in the questionnaire.

Representation of the survey: there were interviewed 1188 MMU managers from whom 1064 or 89.6% were micro enterprises and 124 or 10.4% small size enterprises. According data of the Central Statistical Bureau of Republic of Latvia in 2010 there were 91.1% micro enterprises and 7.2% small enterprises. Most reflected activity fields in the survey were trade – 22.2% of respondents, agriculture, forestry and fishery – 17.6%, other services – 11.9%, professional, scientific and technical services – 11.4%, etc (table 1) which corresponds with data of the Central Statistical Bureau of Republic of Latvia (Number of Companies, CSB, 2012).

Table 1

Comparison of Response Shares of the Survey and CSB Results on Kinds of Activities

Kind of Activity	**Share of Micro and Small Size Enterprises (%)**	
	CSB data*	**MSE survey data**
Trade, car and motorbike reparation	18.2	22.2
Agriculture, forestry and fishery	22.6	17.6
Other services	9.4	11.9
Professional, scientific and technical services	8.9	11.4
Processing industry	5.4	9.5
Construction	5.0	7.1
Transport	4.0	5.2
Information and communication services	2.4	4.8
Health care and social care	3.1	3.6
Hospitality and catering services	2.3	3.6
Electrical power, gas supply, heating and air conditioning	0.3	0.9

*Source: Ināra Kantāne calculations, CSB data and Ināra Kantāne conducted survey (December 2010 – August 2011), sample size n = 1188; * http://data.csb.gov.lv/Dialog/Saveshow.asp* – observed 20.04.2012.

After three times attempt to every possible respondent the response rate was 21.7%. Distribution of respondents in statistical regions: Rīga region – 35.3%, Pierīga – 16.9%, Kurzeme region – 14.2%, Latgale region – 13.1%, Vidzeme region – 11.2%, Zemgale region – 9.3% respondents which corresponds to the data of Central Statistical Bureau of Latvia on distribution of micro and small enterprises by regions (Table 2).

Table 2

Comparison of Shares of CSB and Survey Data by Regions of Latvia

Region	**Share of Micro and Small Size Enterprises (%)**	
	CSB data*	**MSE survey data**
Rīga	38.2	35.3
Pierīga	15.3	16.9
Vidzeme	10.8	11.2
Kurzeme	12.6	14.2
Zemgale	10.4	9.3
Latgale	12.7	13.1

*Source: Ināra Kantāne calculations, CSB data and Ināra Kantāne conducted survey (December 2010 – August 2011), sample size n = 1188; * http://data.csb.gov.lv/Dialog/Saveshow.asp* – observed 20.04.2012.

Survey results and CSB data (population data) differ only by some percent, those differences are not significant for all regions and for all kinds of activities.

Evaluations of micro and small enterprise managers on internet use indicates that many of managers evaluate highly internet use, but still there are managers who are not so fond of internet use, half of the respondents evaluated importance of the internet use by 8 (in 1 – 10 scale) and half of the respondents have evaluated less than 8 (median), but the most often evaluation was 10 – mode (Table 3).

Table 3

Evaluations of MSE Managers on Internet Use

Statistical indicators	**Values of Statistical Indicators**
Number of respondents	1051**
Arithmetic mean	7.22
Standard Error of Arithmetic Mean	0.09
Median	8.00
Mode	10
Standard deviation	3.115
Range	10
Minimum	0
Maximum	10

*Source: Calculations on Ināra Kantāne conducted survey (December 2010 – August 2011), sample size n = 1188; Evaluation scale 1 – 10, where 0 – do not use; 1 –use very seldom; 10 –use very often; **number of replied respondents*

54.7% of MSE managers have indicated that they actively use internet (evaluation 8 – 10), 38.64% managers have indicated very widely, but 6.9%, of MSE managers indicated that they do not use internet (Figure 1).

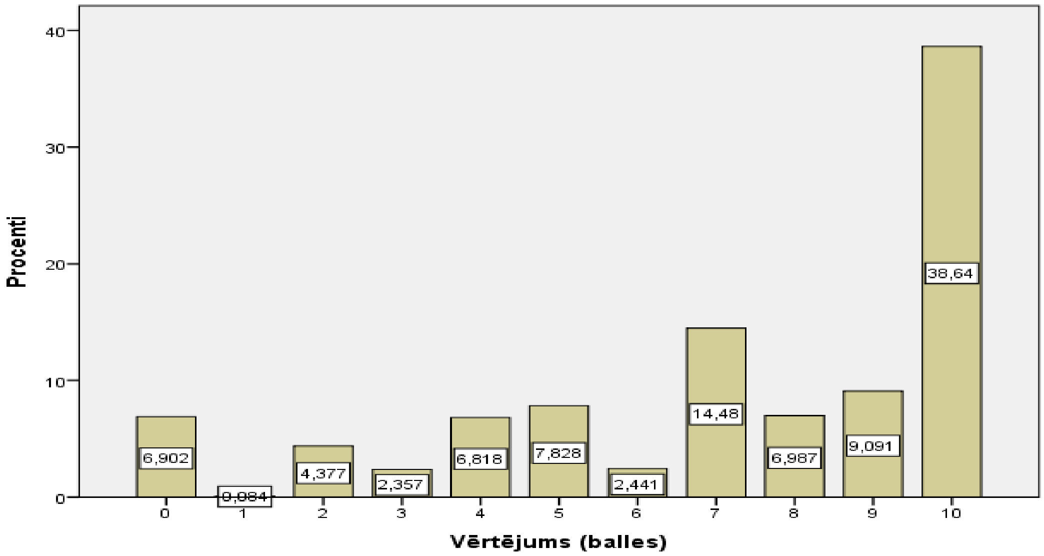


Figure 1. Distribution of Responses on Internet Use in Micro and Small Enterprises

Source: Calculations on Ināra Kantāne conducted survey (December 2010 – August 2011), sample size n = 1188; Evaluation scale 1 – 10, where: 0 – do not use; 1 – use very seldom; 10 – use very often

MSE managers have mentioned that mostly internet they use for communication with state and municipality institutions, clients and other business persons.

MSE managers on question about the company webpage have responded the following: around half of respondents (44.8%) have replied that they have company webpage, 17.53% of managers indicated that company webpage is under construction, but 28.57% of respondents have indicated that they do not have company webpage (Figure 2). Survey results correspond to information on company webpages by Central Statistical Bureau of Republic of Latvia (Information Technologies, CSB, 2012).

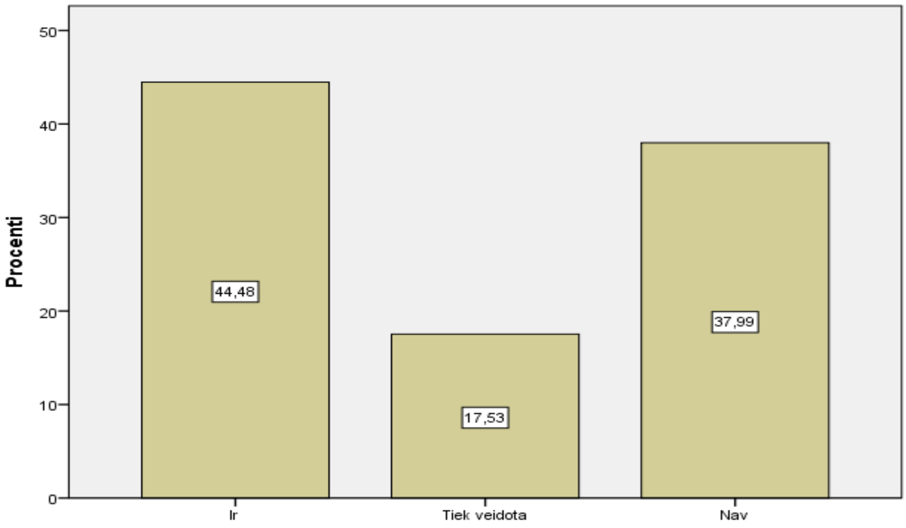


Figure 2. Distribution of Responses on Availability of Company Homepage in MSE

Source: Calculations on Ināra Kantāne conducted survey (December 2010 – August 2011), sample size n = 1188

MSE managers have mentioned that internet webpage can ensure with more precise information for service receivers and, most important it is to provide information about the respective company possibly wider audience – current and future clients.

Obtained results on internet use confirm that 6.9% of the companies do not use internet, but 13.6% of the companies internet use seldom what influences communication with government institutions, use of e-services, obtaining information on state support and other important issues as well as information exchange in the company, communication with clients and suppliers. MSE managers who do not use internet are in great extent excluded from actual business competition.

Conclusions

Micro and small size enterprises in Latvia are not well equipped with the internet, survey conduction via internet is not possible on full extent. To conduct surveys for those companies lacking internet connections or abilities of internet use it is necessary to reserve additional time and other resources to get information from companies not having access to the Internet and having limited abilities of information technologies use.

For different evaluations in the survey evaluation scale 1 – 10 has been acceptable and well understandable for respondents.

References

Black, I.R., Efron, A., Ioannou, C., Rose, J.M. (2005). Designing and Implementing Internet Questionnaires Using Microsoft Excel, *Australasian Marketing Journal (AMJ)*, Volume 13, Issue 2, pp. 61-72.

Chang, M.K., Cheung, W. (2001), Determinants of the intention to use Internet/WWW at work: a confirmatory study, *Information & Management*, Volume 39, Issue 1, pp. 1-14.

Dennis, W.J. (2003). Raising response rates in mail surveys of small business owners: results of an experiment. *Journal of Small Business Management*, Volume 41, Issue 3, pp. 278-295.

van Dijk, J.A.G.M, Peters, O., Ebbers, W. (2008). Explaining the acceptance and use of government Internet services: A multivariate analysis of 2006 survey data in the Netherlands, *Government Information Quarterly*, Volume 25, Issue 3, pp. 379-399.

Dillman, D.A. (2000). Mail and Internet Surveys: The Tailored Design Method. John Wiley & Sons, New York, NY.

Frohlich, M.T. (2002). Techniques for improving response rates in OM survey research. *Journal of Operations Management*, Volume 20, Issue 1, pp. 53-62.

Garratt, A.M., Helgeland, J. Gulbrandsen, P. (2011). Five-point scales outperform 10-point scales in a randomized comparison of item scaling for the Patient Experiences Questionnaire *Journal of Clinical Epidemiology*, Volume 64, Issue 2, pp 200-207.

Ilieva, J., Baron, S., Healey, N.M. (2002). Online surveys in marketing research: pros and cons. International *Journal of Market Research*, Volume 44, Issue 3, pp. 361-382.

Janes, J. Survey construction, *Library Hi Tech*, Volume 17, Issue 3, 1999, pp. 321-325.

Kelly, P. (2000). Questionnaire design, printing, and distribution, *Government Information Quarterly*, Volume 17, Issue 2, pp. 147-159.

Knight, G.A. Entrepreneurship and strategy in the international SME, *Journal of International Management*, Volume 7, Issue 3, 2001, pp. 155-171.

Melnyk, S.A., Page, T.J., Wu, S.J., Burns, L.A. (2012). Would you mind completing this survey: Assessing the state of survey research in supply chain management, *Journal of Purchasing and Supply Management*, Volume 18, Issue 1, pp. 35-45.

Number of Companies, CSB of Latvia – *http://data.csb.gov.lv/Dialog/Saveshow.asp* – 20.04.2012.

Redoli, J., Mompó, R., García-Díez, J., López-Coronado, M. (2008). A model for the assessment and development of Internet-based information and communication services in small and medium enterprises. *Technovation*, Volume 28, Issue 7, pp. 424-435.

Science and Technologies, Information Technologies, Central Statistical Bureau of Republic of Latvia – *http://data.csb.gov.lv/Dialog/Saveshow.asp* – 22.04.2012.

Sivo, S.A., Saunders, C., Chang, Q., Jiang, J.J. (2006). How low should you go? Low response rates and the validity of inference in IS questionnaire research. *Journal of the Association for Information Systems,* Volume 7, Issue 6, pp. 351-414.

Smith, M.H., Smith, D. (2007). Implementing strategically aligned performance measurement in small firms, *International Journal of Production Economics*, Volume 106, Issue 2, pp. 393-408.

Tsatsou, P., (2011). Why Internet use? A quantitative examination of the role of everyday life and Internet policy and regulation, *Technology in Society*, Volume 33, Issues 1-2, pp. 73-83.

Verdegem, P., Verleye, G. (2009). User-centered E-Government in practice: A comprehensive model for measuring user satisfaction, *Government Information Quarterly*, Volume 26, Issue 3, pp. 487-497.

On sample allocation for effective EBLUP estimation of small area totals

Mauno Keto[1]

[1]Mikkeli University of Applied Sciences - Finland, e-mail: mauno.keto@mamk.fi

Abstract

The demand of regional or small area statistics produced from large-scale surveys has raised needs for developing the tools of optimal sample allocation on area level. The concept of optimality can of course be defined in many different ways. Most commonly used allocation methods aim at producing efficient direct areal estimates. What often happens, however, is that due to sparse sampling resources several areas receive little or none observations, and therefore indirect estimation may be necessary. These methods are often based on nested-error regression type model-based estimators. For this reason should areal sample allocation be implemented in such a way that it would lead to efficient estimation in the case of an indirect estimator. The problem has been tried to solve in this research by developing an analytical allocation method based on the main component of MSE in EBLUP estimation. The performance of this method has been tested by comparing it to various other allocations through sample simulations from real data. The effectiveness of each allocation was measured with MSE, CV and certain quality measures (ARE, ARB and RRMSE). Some results are presented here.

Key words: Planning samples sizes of small areas, indirect estimation, optimal allocation, optimization criterion, areas with none observations.

1 Introduction

We plan sampling designs generally for efficient estimation on the population level. However, the same demand of efficiency prevails if one wants to calculate regional or small area statistics from large-scale survey data but now on the level of some subpopulation. Generally, as for basic sampling design, stratified random sampling has been chosen. Strata coincide with areas and the problem is how to allocate stratum-wise fixed sample size *n*.

Optimal allocation has inspired for different solutions during the last decades. Main line has prevailed to find areal allocation giving possibility to calculate direct or model-assisted direct estimators for each area. Some examples from earlier efforts are reported in Rao (2003). Recently published interesting proposition come from Longford (2006), who includes inferential priority index P_d for each area and tries then to find optimality. Another solution comes from Falorsi and Righi (2008). They assume that direct estimators should be model-assisted and their optimal allocation procedure accounts for this possibility with other prior information used in planning sample design.

The next sections describe different approaches to areal allocation, the selected model as additional information, earlier experimental studies, developing optimal sample sizes in one simple situation, the searching of an areal allocation scheme conditional to auxiliary information which includes both auxiliary variables and model for indirect estimation of fixed areal totals, and finally some results of simulations based on different allocations. Indirect or model-based estimation has been chosen because in small area calculations domains with few or none observations are general. The problem of choise of model has been profoundly investigated by Lehtonen *et al* (2003 and 2006). As a model, EBLUP has been chosen because there is a lot of evidence that this model works well in many small area estimation situations.

2 Brief overview of sample allocation methods in stratified sampling

2.1 Allocation methods developed for population and area level

Equal allocation is based only on the number of areas and doesn´t take the characteristics of areas into account at all. The sample size of each area is simply *n/D*, where *n* = overall sample size and *D* = number of areas. Especially large areas with strong variation in variables of interest suffer from the point of view of efficiency and accuracy.

Proportional allocation can be used when larger areas are expected to have higher variance compared with smaller areas. The sample size of area *d* is proportional to the number of observation units (N_d) in that area:

$$n_{d,pro} = f_d n = (N_d / N)n \, .$$

This allocation ensures same proportion for each area in the sample, but does not guarantee efficient estimation, especially for areas in which the response variable has high variance and significantly higher values compared with smaller areas.

Neyman´s allocation which is a special case of optimal allocation is based on sizes of areas and the variances (or standard deviations) of auxiliary variable *x* used in estimation. The *x*-value of each observation unit in the population or at least its variance in each area must be known. Sample size if area *d* is computed as follows:

$$n_{d,opt} = (N_d S_d / \sum_{d=1}^{D} N_d S_d)n \, ,$$

where S_d = standard deviation of auxiliary variable *x* in area *d*. This allocation favours large areas with high variance. Differences in sample sizes between areas can be significantly large compared with for example proportional allocation.

Power (Bankier) allocation is based more on internal characteristics of areas compared with previously mentioned methods. It is recommended to be used in a research where the population contains many small areas and reliable estimates must be produced for each area. Formula for ample size for area *d* is

$$n_{d,pow} = (X_d^a CV(x)_d) \sum_{d=1}^{D} X_d^a CV(x)_d)n \, ,$$

where X_d means the sum of values of auxiliary variable *(x)* and *CV(x)*$_d$ is the coefficient of variation of variable *x* in area *d*. Exponent *a* is a certain power value which can be used to regulate the significance of variable *x*. Values ½ and ⅓ are often recommended for *a*.

2.2 Approaches to optimal allocation based on certain assumptions and criteria

Some articles concerning optimal sample allocation in stratified sampling have been published during last years. Two of them are shortly referred in this paper.

Longford (2006) has introduced a method to calculate optimal sample sizes for areas when minimizing the weighted sum of sampling variances in direct estimation as a function of sample sizes n_d. Minimum can be found only under simple assumptions. The weights are called inferential priorities and their values can be changed and thus areas can be given different significance.

Falorsi and Righi (2008) have used a sampling strategy to determine sample sizes for domains that is based on balanced sampling when domains (areas) have first been divided into non-overlapping partitions in many different ways. The main goal was to produce sample sizes which guarantee the sampling errors of domain estimates to be lower than given limits. Optimization of inclusion probabilities is a crucial part of the method. Many different estimators were tested through simulations. The authors finally ended up to recommend a GREG-type model-assisted estimator.

3 Model and earlier experiments related to use of model

3.1 Model

The model used in this research is a nested-error regression, basic unit level model

$$y_{dk} = \mathbf{x}'_{dk}\boldsymbol{\beta} + v_d + e_{dk};\ \ k = 1,\dots,N_d; d = 1,\dots,D \qquad (1)$$

which is a special case of well-known general mixed linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zv} + \mathbf{e}, \qquad (2)$$

where **y** is n×1 vector of sample observations, **X** and **Z** are known n×p and n×h matrices of full rank, and **v** and **e** are independently distributed with means **0** and covariance matrices **G** and **R** depending on some variance parameters **δ** = $(\delta_1,\dots,\delta_q)'$. Furthermore, Var(**y**) = **V** = **R** + **ZGZ′** is the variance-covariance matrix of **y**. In model (1) y_{dk} is the kth value in area *d* for outcome variable (y), $\mathbf{x}_{dk}$ is the vector of auxiliary variables (x) in area *d*, v_d is the latent random effect of area *d* (*d* = 1,…,*D*) in the model and is estimated from the observations, and e_{dk} is a random error. Random effects v_d and random errors e_{dk} are assumed to be independent of each other and distributed with mean zero and variances σ_v^2 and σ_e^2 (not necessarily normally). Regression coefficients $\boldsymbol{\beta}$ are estimated from the observations.

Regression coefficients $\boldsymbol{\beta}$ and area effects **v** are estimated from the observations as follows:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad \hat{\mathbf{v}} = \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad (3)$$

EBLUP (Empirical Best Linear Unbiased Predictor) estimator for area total Y_d is the sum of sample observations and predicted values of non-sampled observations of variable *y* as given in Rao (2003):

$$\hat{Y}_{d,EBLUP} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \hat{y}_{dk} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \mathbf{x}'_{dk}\hat{\boldsymbol{\beta}} + (N_d - n_d)\hat{v}_d \qquad (4)$$

The MSE of estimator $\hat{Y}_{d,EBLUP}$ is the sum of its variance and squared bias:

$$MSE(\hat{Y}_{d,Eblup}) = E(\hat{Y}_{d,Eblup} - Y_d)^2 = Var(\hat{Y}_{d,Eblup}) + (\hat{Y}_{d,Eblup} - Y_d)^2 \qquad (5)$$

An estimator of MSE approximation in the case of finite populations is given in Rao (2003):

$$mse(\hat{Y}_{d,EBLUP}) = g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{2d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2g_{3d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{4d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) \qquad (6)$$

The first and most important component which is important later in this presentation is given by

$$g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = (N_d - n_d)^2(1-\gamma_d)\hat{\sigma}_v^2, \text{ where } \gamma_d = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 n_d^{-1}) = \hat{\sigma}_v^2 / (n_d\hat{\sigma}_v^2 + \hat{\sigma}_e^2). \qquad (7)$$

In addition, we define a specific common intrastratum or intra-area correlation

$$\hat{\rho} = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_e^2) = 1/(1 + \hat{\sigma}_e^2 / \hat{\sigma}_v^2), \qquad (8)$$

which measures the proportion of variation between areas and total variation (value between zero and one).

The model (1) is used as additional information when searching for optimal allocation in an analytical way which is based on the structure of model and its MSE.

3.2 Experimental allocation as the first approach under the model

The first approach to allocation problem was “Experimental allocation” which Keto and Pahkinen (2009) have introduced in a conference paper. The idea was shortly following: 1500 SRSWOR-samples were drawn from a population of 400 Finnish municipalities in 19 provinces which served also for areas and strata. Response variable (y) was number of unemployed people and one auxiliary variable was used. Total number of unemployed in each province (area) was estimated by using model (1) and EBLUP estimation. In the first phase MSE, CV and certain quality measures (ARE, ARB etc.) were produced for each area in every sample,and finally their means were computed in every sample. After this, samples were arranged in ascending order according to means. The following figure shows an example of MSE means. Sample sizes for the second phase were determined by using quartiles of these distributions. In the second phase the competence of this “experimental” allocation was compared with three other allocation methods.

Figure 1: Distribution of areal sample sizes in 20 “best” samples for MSE means (boxplot).

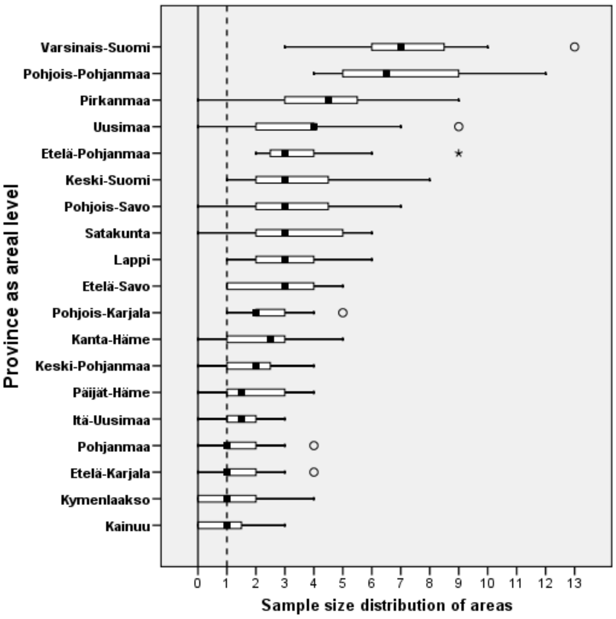


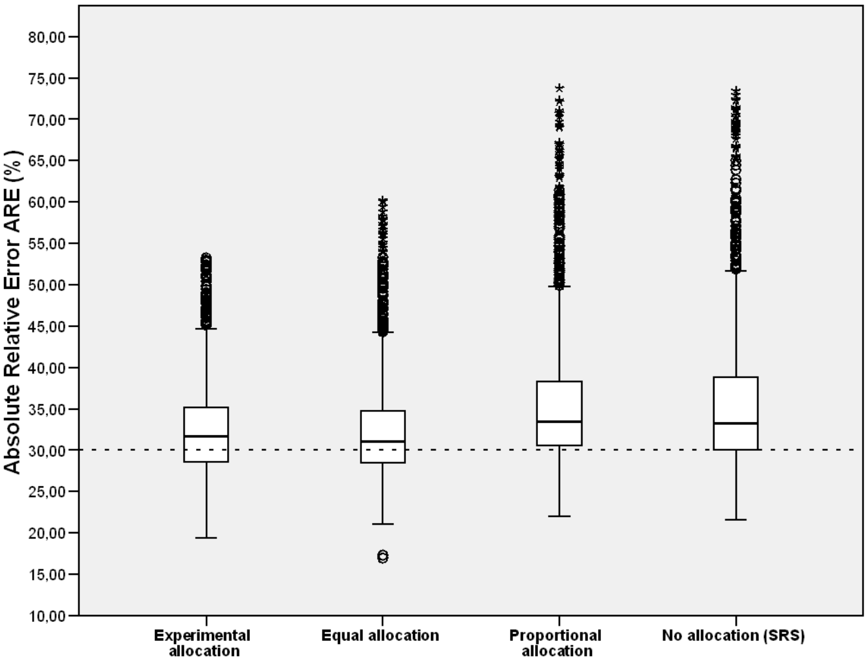
Table 1 shows final sample sizes of areas in each compared allocation. It is worth noticing that in experimental allocation as many as 7 areas have sample size zero.

Table 1: Areal sample sizes in different allocation schemes

Province	Size of area	Not allocated	Proportional	Equal	Experimental
Uusimaa	24		4	3	4
Varsinais-Suomi	53		7	3	7
Itä-Uusimaa	10		1	3	0
Satakunta	25	S	4	3	5
Kanta-Häme	16	R	2	3	0
Pirkanmaa	28	S	4	3	5
Päijät-Häme	12	-	2	3	0
Kymenlaakso	12	s	2	3	0
Etelä-Karjala	12	a	2	3	0
Etelä-Savo	18	m	3	3	3
Pohjois-Savo	23	p	3	3	4
Pohjois-Karjala	16	l	2	3	3
Keski-Suomi	28	e	4	3	5
Etelä-Pohjanmaa	26	s	4	3	6
Pohjanmaa	17		2	3	4
Keski-Pohjanmaa	12		2	3	0
Pohjois-Pohjanmaa	38		5	3	6
Kainuu	9		1	3	0
Lappi	21		3	3	5
TOTAL:	**400**	**57**	**57**	**57**	**57**

1 500 SRSWOR samples were simulated from population by using each of these allocations (4 times 1500 samples) and same statistics (MSE, CV) and quality measures (ARE, ARB etc.) were computed. Figure 2 shows the distributions of average absolute relative error (ARE, see appendix) in the samples.

Figure 2: Distributions of 95 % of ARE values of samples



Of course experimental allocation method cannot be used to prove the better performance of tested allocation, but it can show the topics to focus on in future research concerning effective sample allocation.

4 Example of finding optimal sample sizes in a simple situation

Let us assume that for each observation unit in the population there exists value x_{dk} of auxiliary variable *x*, and this value is correlated with value y_{dk} of response variable *y*, which means that following equation holds between these values: $y_{dk} = f(x_{dk}) + e_{dk}$, where e_{dk} means additive residuals. The variance of response variable can be computed with variance of explanatory variable. If variances are finite, we get equation

$$V(y_{dk}) = E(V(y_{dk}|x_{dk})) + V(E(y_{dk}|x_{dk})). \tag{9}$$

Let us consider a unit-level linear regression model

$$y_{dk} = \alpha + \beta x_{dk} + e_{dk}, e_{dk} \sim N(0,\sigma^2),$$

where α and β are least-squares regression coefficients estimated from the sample. By using rule (9) we get an estimator for variance of y_{dk} as follows: $\hat{V}(y_{dk}) = \beta^2 V(x_{dk}) + \sigma^2$. The estimator of the mean of response variable *y* in area *d* is given by expression

$$\hat{V}(\bar{y}_d) = (1 - n_d/N_d)(\beta^2 V(x_{dk}) + \sigma^2)/n_d = (1/n_d - 1/N_d)(\beta^2 V(x_{dk}) + \sigma^2)$$

that contains finite population correction. The mean of areal variances can be written as

$$(1/D)\sum_{d=1}^{D}(1/n_d - 1/N_d)(\beta^2 V(x_{dk}) + \sigma^2) = (1/D)\sum_{d=1}^{D}(1/n_d)(\beta^2 V(x_{dk}) + \sigma^2) - (1/D)\sum_{d=1}^{D}(1/N_d)(\beta^2 V(x_{dk}) + \sigma^2).$$

This expression can be minimized as a function of sample sizes $n_1,...,n_D$ by using a well-known Lagrange´s method under constraint $\sum_d n_d = n$. The result for optimal sample size is (derivation can be proved)

$$n_d = (\sqrt{1+(\beta^2/\sigma^2)V(x_{dk})} / \sum_d \sqrt{1+(\beta^2/\sigma^2)V(x_{dk})}) \times n\ .$$

If quantity $(\beta^2/\sigma^2)V(x_{dk})$ is large enough, value one is negligible, so we get an approximate value for n_d :

$$n_d \approx (\sqrt{(\beta^2/\sigma^2)V(x_{dk})} / \sum_d \sqrt{(\beta^2/\sigma^2)V(x_{dk})}) \times n = ((\beta/\sigma)S(x_{dk}) / (\beta/\sigma)\sum_d S(x_{dk})) \times n$$
$$= (S(x_{dk}) / \sum_d S(x_{dk})) \times n\ .$$

Optimal sample size is approximately proportional to areal standard deviation of covariate *x*. The same result can be obtained also without finite population correction. But the result was obtained through minimization of mean of variances. Some areal variances may remain considerably high. The purpose of this example is to show that optimality or at least approximate optimality can be reached under sufficient simple assumptions.

5 Searching for optimal allocation analytically under selected model

Let us return to model (1). It contains fixed part (regression) and a part containing random area effects. A random sample of *n* sampling units is selected from *D* areas, and sampling method is SRSWOR inside strata. Furthermore, $\sum_d n_d = n$. EBLUP estimation produces normally estimates $\hat{Y}_d$ for areal totals of response variable *(y)* according to (4) and MSE approximations (6). Also CV´s of areal estimates can be computed.

First we try to select the sample from strata so that we can minimize the arithmetic mean of areal MSE´s as a function of sample sizes $n_1, n_2,...,n_D$. But because MSE has a very complex expression, the minimization is not possible. We turn attention to the first part of MSE, g_{1d} (7). According to Nissinen (2009), this term has contributed 85 – 95 % of total MSE in many surveys, but this requires sufficient variation between areas.

Criterion for optimal allocation is now minimizing mean of areal g_{1d} values

$$1/D\sum_{d=1}^{D} g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = 1/D\sum_{d=1}^{D}(N_d - n_d)^2(1/\hat{\sigma}_e^2 \times n_d + 1/\hat{\sigma}_v^2)^{-1}\ . \qquad (10)$$

as a function of sample sizes subject to constraint $\sum_d n_d = n$. We use method of Lagrange multipliers to solve each n_d. Derivation process is not shown here because of its length, but the result can be proved. The expression for sample size of area *d* is

$$n_{d,opt} = -\hat{\sigma}_e^2 / \hat{\sigma}_v^2 + \frac{(N_d + \hat{\sigma}_e^2 / \hat{\sigma}_v^2)(n + D(\hat{\sigma}_e^2 / \hat{\sigma}_v^2))}{N + D(\hat{\sigma}_e^2 / \hat{\sigma}_v^2)}\ . \qquad (11)$$

This expression contains ratio $\delta = \hat{\sigma}_e^2 / \hat{\sigma}_v^2$ of variance components which depends on the sample. We have earlier defined a specific intra-area correlation $\hat{\rho}$ in (8), and ratio δ is given now by expression

$$\delta = 1/\hat{\rho} - 1 . \tag{12}$$

Expression (10) can now be written in the form

$$n_d = -\delta + \frac{(N_d + \delta)(n + \delta D)}{N + \delta D} = \frac{N_d n - (N - N_d D - n)\delta}{N + D\delta} = \frac{N_d n - (N - N_d D - n)(1/\hat{\rho} - 1)}{N + D(1/\hat{\rho} - 1)} . \tag{13}$$

Some conclusions can be made immediately when examining expressions (11) or (13): 1) expression is meaningful when variance component $\hat{\sigma}_v^2 > 0$, 2) because $n + \delta D < N + \delta D$ it follows that $n_d < N_d$, 3) value of n_d depends on the ratio of variance components but not directly on the values of variance components, and 4) if all total variation consists only of variation between areas ($\delta = 0$), final result would be proportional allocation. More precise examination reveals that sample size n_d can become negative in certain situations: area *d* is small, overall sample size *n* is small and total variation is mostly within areas.

Because ratio $\delta = \hat{\sigma}_e^2 / \hat{\sigma}_v^2$ cannot be used to compute sample sizes, we have to replace it with a corresponding value that can be obtained from auxiliary variable. Because response variable *y* and auxiliary variable *x* are correlated, we assume that the variation of auxiliary variable transfers to the sample. We use so called homogeneity measure which is related to cluster sampling and which is presented for example by Särndal *et al* (1992). If the clusters have different sizes then homogeneity measure is given by expression

$$R_a^2 = 1 - R^2 = 1 - \frac{MSW}{S^2} , \tag{14}$$

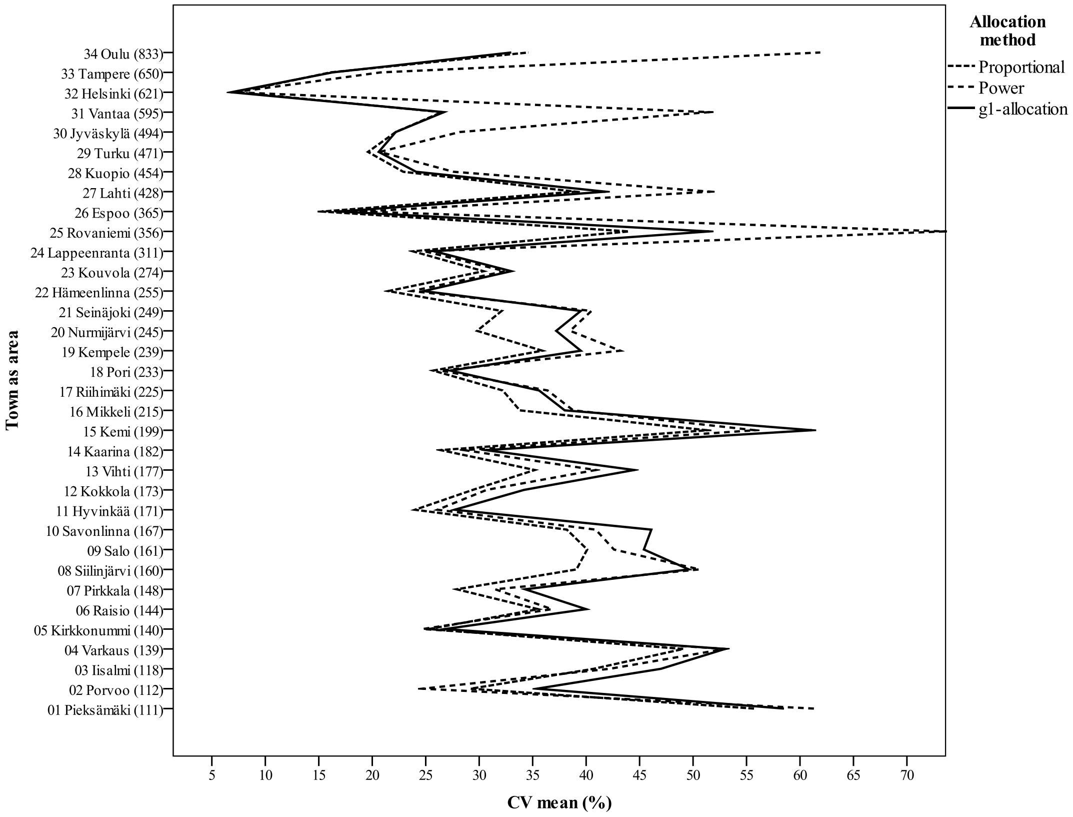
where R^2 means coefficient of determination (familiar from regression analysis), MSW is mean sum of squares within clusters (in this case strata), and S^2 is the variance of auxiliary variable. When substituting homogeneity measure (14) for intra-area correlation ($\hat{\rho}$) in (12) the sample sizes for areas can be calculated.

This method was applied to a case in which the population consists of 9 815 apartments for sale in 34 Finnish towns. The data were collected from an internet source in 2011.The size of smallest area was 111 (apartments, sampling units) and size of largest area was 833. Variable *(y)* measures the price of apartment (1 000 €) and auxiliary variable *(x)* measures size (m^2). Variables are correlated. Overall sample size *(n)* was very low 102 (3 times 34). Value of homogeneity measure for auxiliary variable *(x)* for computing sample sizes is 0.33 which means that variation between areas was high. Some compromises had to be made (for ex. negative sample sizes were turned to zero). Final areal sample sizes varied from zero (three areas) to 12.

To test the performance of developed method (we call it "g1-allocation") it was compared with five other allocations: SRSWOR, equal, proportional, optimal (Neyman) and power allocation. 1 500 samples were simulated for each method (6 times 1 500 samples). Sampling method was SRSWOR inside strata, except for first where allocation was not used. MSE and CV plus certain quality measures to discover accuracy and bias were calculated for areas in each sample, as well as the areal means of these statistics and measures.

Figure 3 presents areal CV means for calculated of 1500 samples for three different allocations (proportional, power and g1-allocation) which take areal characteristics into account. Optimal (Neyman), equal and SRSWOR alternatives are not presented. One can notice that g1-allocation had good performance in large areas, but fairly poor performance on some smaller areas. But what must be mentioned is the fact that three smallest areas were non-sampled areas in g1-allocation. Presented results bring hopefully new aspects to consider in model-based small area estimation. More investigation of the influence of areas is needed.

Figure 3: Areal CV means computed from 1 500 samples (sizes of areas inside brackets)



Appendix: formulas

Coefficient of variation (CV) for estimate of area total Y_d in EBLUP estimation:

$$CV(\hat{Y}_{d,EBLUP})\% = 100\times(\sqrt{mse(\hat{Y}_{d,EBLUP})}\,/\,\hat{Y}_{d,EBLUP})$$

Average absolute relative error (ARE) in one sample for estimate of area total Y_d in EBLUP estimation:

$$ARE\ \% = 100\times(1/D)\sum_{d=1}^{D}\left|\hat{Y}_{d,EBLUP} - Y_d\right|/\,Y_d$$

where D = number of estimated areas.

References

Falorsi, P.D. and Righi, P. (2008). A balanced sampling approach for multi-way stratification for small area estimation. *Survey Methodology* **34,** 223-234.

Keto, M. and Pahkinen, E. (2009). On sample allocation for effective EBLUP estimation of small area totals – "Experimental Allocation". In: J. Wywial and W. Gamrot (eds.). (2010). *Survey Sampling Methods in Economic and Social Research*. Katowice: Katowice University of Economics.

Khan, M.G.M., Maiti, T. and Ahsan, M.J. (2010). An Optimal Multivariate Stratified Sampling Design

Using Auxiliary Information: An Integer Solution Using Goal Programming Approach. *Journal of Official Statistics* **26,** 695-708.

Lehtonen, R., Myrskylä, M., Särndal, C.-E. and Veijanen, A. (2006). The role of models in model-assisted and model-dependent estimation for domains and small areas. *Working paper, BNU Workshop,* Ventspils, Latvia, August 2006.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The Effect of Model Choice in Estimation for Domains, Including Small Domains. *Survey Methodology* **29,** 33-44.

Longford, N. T. (2006). Sample Size Calculation for Small-Area Estimation. *Survey Methodology* **32,** 87 - 96.

Nissinen, K. (2009). *Small Area Estimation With Linear Mixed Models From Unit-Level Panel and Rotating Panel Data.* University of Jyväskylä, Department of Mathematics and Statistics, Report **117**. (Dissertation).

Rao, J. N. K. (2003). *Small Area Estimation*. Hobogen, New Jersey: Wiley.

Särndal, C-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Verlag.

Organisation of Surveys on Attitudes Towards Hydrogen as Energy Carrier

Biruta Sloka[1] **and Jānis Kleperis**[2]
Justs Dimants[3] **and Ilze Dimanta**[4]
Māra Gudakovska, Jānis Kleperis Jr. and Pēteris Tora
[1]University of Latvia, e-mail: Biruta.Sloka@lu.lv
[2]University of Latvia, Institute of Solid State Physics, e-mail: kleperis@latnet.lv
[3]University of Latvia, e-mail: Justs.Dimants@lu.lv
[4]University of Latvia, e-mail: Ilze.Dimanta@lu.lv

Abstract

Use of renewable energy resources, including hydrogen energy is on agenda for energy developers. Popularity of renewable energy is growing constantly. There are lots of successful projects and more often companies and different societies start to implement renewable energy projects to manage efficient financial resource spending as well as reduce the impact of energy suppliers. Lots of good practice examples are examined and developed world wide, including operation of university campus, public transport, operation of villages, etc. Paper examines surveys and their organisation for the readiness of acceptance of renewable energy resources and in this case – hydrogen for energy supply of Academic Centre of University of Latvia. In the survey were questions on respondent's, attitudes, behaviour, some environmental knowledge as well as information on socio-economic characteristics of respondents, including, questions about the hydrogen energy acceptance, scientific value and safety issues. The main conclusions are that developed survey organisation can be used also for research in other groups of segments and the main survey results shows acceptance for the hydrogen as energy source. Methods used for analysis: scientific publications research, evaluation of practical knowledge transfer and public opinion examination tools and marketing tools application evaluations using questionnaires. For data processing and analysis indicators of central tendency or location and variability, crosstabulations were used.

Keywords: surveys on attitudes, knowledge evaluation, public acceptance, hydrogen energy

1 Introduction and theoretical background

Already for several decades researchers worldwide work on hydrogen as energy carrier, on opinion research about those issues – the results are discussed in solid international scientific conferences and published in numerous scientific publications. Different countries have different approach and different attitude of public and politicians as well as implementation of the findings in everyday life: on acceptance of hydrogen technologies lot of research is done in Germany (Altmann, *et al.* 2012*)*, in Australia (Dicks, *et al.* 2004), in Wales (Cherryman, *et al.* 2005), in the Netherlands (Zachariah – Wolff, *et al. 2004),* in China (Cropper, 2002a), in India (Cropper, 2002b), in Norway (Bak, 2003), in USA (Bak, 2004) and (Schmoyer, *et al.* 2004), in Brazil (Hotza, *et al.* 2008), in London, Teeside ans Wales (Ricci, *et al. 2006),* in Lithuania (Milciuviene, *et al. 2006),* in some extent also in Latvia (Dimants, *et al,* 2012), in Iceland many discussions are realised and already implemented – hydrogen as future hydrogen economy (Aranson, *et al.* 2000), on public acceptance (Dressner, *et al.* 2007) and (Wilsdon, *et al. 2004),* on hydrogen production (Turner, *et al.* 2008 and by UNEP, 2006), as energy carrier ((Wietschel, *et al.* 2007) and (Zhang, 2010)), on hydrogen technologies (Hagen, 2003) and (Waegel, *et al. 2006),* on hydrogen production from waste (Rabah, *et al.* 1989), on issues in the public perceptions of risk (Flynn, 2004). Different sources for hydrogen production are evaluated by political, economical and environmental aspects (Balat, *et al.* 2009), (Meisen, 1996), (Ricci, *et al.* 2008 and 2007 and 2006), (Wilk, *et al.* 2007), on fuel cells (Cropper, *et al.* 2004), on transition to renewable energy systems with hydrogen as an energy carrier (Barbir, 2009), forecasts, scenarious as well as visions are evaluated by McDowal and Eames (McDowall, *et al.* 2006, 2007). Hydrogen futures toward a sustainable energy system is covered in several research works (Dunn, 2001), as well in European Commission, for transport (Farell, *et al.* 2003), and (Bellaby, *et al.* 2007), (Li, *et al.* 2010) for public buses in Stokholm (Haraldsson, *et al.* 2006), several aspects on hydrogen vehicles (O'Garra, 2012, 2007 and 2005), on hydrogen fueling stations (Fuel Cells, 2000). A global survey of hydrogen energy research, development and policy are in depth researched already many years ago (Solomon, *et al.* 2006).

The term "*hydrogen economy*" was formulated in 1970 by the 20th century remarkable electrochemist John O'Mara Bockris (Bockris, 2002) as an alternative to oil and coal-based economy of today. The non-renewable energy resources of oil, natural gas, coal on the Earth are limited and not restored quickly enough to compensate the growing consumption year from year. But hydrogen, although does not occur in the free form on Earth, can be obtained using renewable energy (wind, sun, water, geothermal) and renewable resources (biomass, water). Hydrogen as a fuel can be used for transport, and production of heat and electricity; the hydrogen combustion (both directly in internal combustion engines, boilers and chemically in fuel cells) does not pollute the environment with carbon and its compounds (soot, hydrocarbons, carbon monoxide CO, carbon dioxide CO_2. In addition, if fossil energy resources on Earth are not everywhere, and the battle for ownership of the deposits is related to the cruel wars of all time, even today, when renewable energy and renewable resources are to be acquired in almost every country in the world. Therefore, a wide transition to hydrogen as an energy carrier and fuel, or *Hydrogen Economy* marks the start of a new era, characterized by greater energy independence and less environmental pollution. For hydrogen, there are many myths, and most important of which related to hydrogen as an unsafe and even dangerous fuel. Recent public interest in hydrogen has elicited a great deal of conflicting, confusing, and often ill-informed commentaries, therefore peer-reviewed white paper for both lay and technical readers was published in the United States by Amory Lovins (Lovins, 2003), documented primer on basic hydrogen facts, weighs competing opinions, and corrects twenty widespread misconceptions. United States Department of Energy in 2001 postulated key components while transition to Hydrogen Economy is necessary (United States Department of Energy, 2002):

- Hydrogen is "The Freedom Fuel";
- Hydrogen provides independence and an environmental choice;
- Hydrogen solves foreign oil dependency and improves the environment:
- Hydrogen is everywhere—"it's right in our backyard";
- A hydrogen economy includes other fuels and
- Hydrogen—it works (it is an ongoing business today);
- Hydrogen is safe;
- Hydrogen is a long-term energy solution;
- Hydrogen is the "man on the moon" equivalent for this generation.

With hydrogen technology information dissemination in Latvia works the Latvian Hydrogen Association (www.h2lv.eu) whose active members are students – authors of this study.There are several questionnaires used in research, our choice was evaluation scale 1 – 10, as it is more and more used for attitude evaluations.

2 Main results

University of Latvia as Organization should choose economically viable long term energy consumption by promoting sustainable development as well as science development. That is possible, renewable energy technologies will be integrated in the campus energy system. The faculties of natural sciences imply implementing innovative building technologies to provide with electricity Academic Centre of Natural Sciences, University of Latvia (include Biology, Chemistry, Geography and Earth Sciences – research laboratories, lecture-rooms, professor rooms etc., 200 researchers and professors, 2000 students. Social-economical survey via questionnaire was performed in February and March, 2012 to explore readiness of the society to use renewable technologies in the University of Latvia campus. Respondents are related to University of Latvia (students, professors, researchers, and possible future students, etc.). Faculties intended to locate in Academic Centre of Natural Sciences participated in the survey. In the survey were questions on respondent's environmental knowledge, attitudes, behaviour as well as information on socio-economic characteristics of respondents. Including, questions about the project acceptance, scientific value and safety issues. Some descriptive statistics (arithmetic mean, mode, median and indicators of variability) on question about information on knowledge about hydrogen as energy resource are reflected in table 1.

Table1. Main statistical indicators of responses on the question "*I am fully informed for hydrogen usability as energy resource"*

Indicators		Values
N	Valid	364
	Missing	0
Mean		6,57
Std. Error of Mean		0,140
Median		7,00
Mode		10
Std. Deviation		2,678
Variance		7,171
Range		9
Minimum		1
Maximum		10

Source: Survey performed by authors in March 2012, n=364
Evaluation scale 0-10, where 0 – do not have information about issue, 1 fully disagree, 10 fully agree

As the survey results show (table 1), most of the respondents are very positive (with surprisingly high evaluations) for respondent knowledge level on hydrogen usability as energy resource has been evaluated above average (arithmetic mean = 6,57, with rather low variability – standard error of mean = 0,140, most of evaluations got the highest evaluation – 10, it is represented by mode (Mo = 10), half of respondents gave evaluation 7 or less, half of respondents gave evaluations at least 7 – it is characterised by median (Me = 7,00). It can be concluded that in average academia and students demonstrated positive attitude towards hydrogen energy and demonstrated good knowledge level about hydrogen technologies and are willing to accept and support technology implementation in University of Latvia Academic Centre of Natural Sciences (more in Dimants, *et al.* 2012*)*. For almost all statements most chosen evaluation was the highest – 10, characterised by mode, except for the statement "I am positively convinced for hydrogen energy safety", where the modal evaluation was 5. For this statement the full range of responses were covered (except 0, it means that all respondents had information on analysed issues and expressed their attitude. Table 2 reflects distribution of the answers by faculty for statement: *I am fully informed for hydrogen usability as energy resource.*

Table 2. Distribution of answers for statement: *I am fully informed for hydrogen usability as energy resource* by faculty

	Faculty represented					Total
	Faculty of Biology	Faculty of Physics and Mathematics	Faculty of Geography and Earth Sciences	Faculty of Chemistry	Riga Technical University	
1	6	0	0	3	0	9
2	3	2	3	5	0	13
3	4	1	8	6	0	19
4	8	2	6	4	0	20
5	10	5	12	8	0	35
6	7	0	8	7	1	23
7	14	6	5	9	0	34
8	7	9	6	11	0	33
9	5	13	4	9	0	31
10	14	10	5	14	0	43
Total	78	48	57	76	1	260

Source: Survey performed by authors in March 2012, n=260
Evaluation scale 0-10, where 0 – do not have information about issue, 1 fully disagree, 10 fully agree

Data of table 2 indicates that most informed on hydrogen usability as energy resource are in Faculty of Physics and Mathematics, less informed are in Faculty of Biology. Table 3 reflects distribution of the answers by gender for statement: *I am fully informed for hydrogen usability as energy resource*.

Table 3. Distribution of answers for statement: *I am fully informed for hydrogen usability as energy resource* by gender

	Gender				Total	
	Female	Share (%)	Male	Share (%)	Number	Share (%)
1	13	6,84	4	2,31	17	4,68
2	12	6,32	4	2,31	16	4,41
3	21	11,06	5	2,89	26	7,16
4	15	7,89	12	6,94	27	7,43
5	22	11,58	18	10,40	40	11,02
6	18	9,47	15	8,67	33	9,09
7	28	14,74	22	12,72	50	13,77
8	18	9,47	27	15,61	45	12,40
9	18	9,47	27	15,61	45	12,40
10	25	13,16	39	22,54	64	17,64
Total	190	100	173	100	363	100

Source: Survey performed by authors in March 2012, n=364
Evaluation scale 0-10, where 0 – do not have information about issue, 1 fully disagree, 10 fully agree

Data of table 3 indicates that male persons are much more informed on hydrogen as energy resource.

3 Conclusions

As the survey results show, most of the respondents are very positive (with surprisingly high evaluations) for respondent knowledge level on hydrogen usability as energy resource. It can be concluded that in average academia and students demonstrated positive attitude towards hydrogen energy and demonstrated good knowledge level about hydrogen technologies and are willing to accept and support technology implementation in University of Latvia Academic Centre of Natural Sciences. Also results indicate that most informed on hydrogen usability as energy resource are in Faculty of Physics and Mathematics, less informed are in Faculty of Biology. Interesting conclusion is that male persons are much more informed on hydrogen as energy resource.

References

Altmann M, Graesel C. The acceptance of hydrogen technologies. Available from: http://www.HyWeb.de/ accepth2; (20.06.2012).

Arnason, B., Sigfusson, T.I. (2000). Iceland—a future hydrogen economy. *International Journal of Hydrogen Energy*. 25 (5). 389–394.

Bak, P.E. (2003). Norway builds hydrogen highway from Stavanger to Oslo. *H2Carsbiz,* 26 November 2003.

Bak, P.E. (2004). Ford and BP take hydrogen and fuel cell cars to Michigan, Florida and California. *H2Carsbiz*, 27 April 2004.

Balat, M., Balat, M. (2009). Political, economic and environmental impacts of biomass-based hydrogen, *International Journal of Hydrogen Energy,* 34, 3589-3603.

Barbir, F. (2009). Transition to renewable energy systems with hydrogen as an energy carrier. *Energy,* 34. 308-312.

Bellaby P, Upham P. (2007). Public Engagement with hydrogen infrastructures in transport. Report for the Department for Transport. DfT Horizon Research Programme. Contract No. PPRO 4/54/2.

Bockris JO'M. (2002) The origin of ideas on a hydrogen economy and its solution to the decay of the environment. *International Journal of Hydrogen Energy* **27**, pp. 731–740.

Cherryman, S., King, S., Hawkes, F.R., Dinsdale, R., Hawkes, D.L. (2005). *Public attitudes towards the use of hydrogen energy in Wales*. University of Glamorgan.

Cropper, M. (2002a). Fuel cells in China: a fuel cell market survey. *Fuel Cell Today*, 21 June 2002.

Cropper, M. (2002b). Fuel cells in India: a survey of current developments. *Fuel Cell Today*, 11 December 2002.

Cropper, M., Geiger, S., Jollie, D. (2004). Fuel cells: a survey of current developments. *Journal of Power Sources*. 131 (1–2). 57–61.

Dicks, A.L., Diniz da Costa, J.A., Simpson, A., McLellan, B. (2004). Fuel cells, hydrogen and energy supply in Australia. *Journal of Power Sources*. 131. 1-12.

Dimants, J., Sloka, B., Kleperis, J., Dimanta, I., Gudakovska, M., Kleperis J. Jr., Tora, P. (2012) Opportunities for hydrogen marketing – public opinion analysis, *Proceedings of International Conference "New Challenges in Economic and Business Development – 2012",* 131-140.

Dresner, S., Tomei, J. (2007). Public acceptability of hydrogen: lessons for policies and institutions. UKSHEC Social Science Working Paper 31. Policy Studies Institute. Available from:, http://www.psi.org.uk/ukshec/pdf/31_Public%20Acceptability_Lessons.pdf (27.06.2012).

McDowall, W., Eames, M. (2006). Forecasts, scenarios, visions, backcasts and roadmaps to the hydrogen economy: a review of the hydrogen futures literature. *Energy Policy.* 34(11). 1236–1250.

McDowall, W., Eames, M. (2007). Towards a sustainable hydrogen economy: Amulti-criteria sustainability appraisal of competing hydrogen futures. *International Journal of Hydrogen Energy.* 32(18). 4611–4626.

Dunn, S. (2001). Hydrogen futures: toward a sustainable energy system. *Worldwatch Paper 157,* Worldwatch Institute, Washington, DC.

Farrell, A.E., Keith, D.W., Corbett, J.J. (2003). A strategy for introducing hydrogen into transportation. *Energy Policy.* 31 (13). 1357–1367

Flynn. R. (2004). Knowing the Unknown: issues in the public perceptions of risk. (2004). Paper presented at the BSA Medical Sociology 36th Annual Conference, University of York, UK, 16–18 September 2004. Available from: http://www.psi.org.uk/ukshec/pdf/York%202004.pdf (03.06.2012).

Fuel Cells (2000). Worldwide hydrogen fueling stations. Available at:

http://www.fuelcells.org/h2fuelingstations.pdf (20.05.2012).

O’Garra T. (2012). Comparative Analysis of the impact of the hydrogen bus trials on public awareness, attitudes and preferences: a comparative study of four cities. Accept H2 Full Analysis Report. Available from: http://www.accepth2.com; (16.06.2012)

O’Garra, T., Mourato, S., Pearson, P. (2005). Analysing awareness and acceptability of hydrogen vehicles: a London case study. *Int JHydrogen Energy*. 30(6). 649–59.

O’Garra ,T., Mourato, S., Pearson, P. (2007). Public acceptability of hydrogen fuel cell transport and associated refuelling infrastructure. In: Flynn R, Bellaby P, editors. Risk and the public acceptance of new technologies. Basingstoke:Palgrave Macmillan.

Hagen, E.F. (2003). Realizing the hydrogen future: the International Energy Agency’s efforts to advance hydrogen energy technologies. *International Journal of Hydrogen Energy*. 28 (6). 601–607.

Haraldsson, K., Folkesson, A., Saxe, M., Alvfors, P. (2006). A first report on the attitude towards hydrogen fuel cell buses in Stockholm. *International Journal of Hydrogen Energy*. 31(3). 317–325.

Hotza, D., Diniz da Costa, J.C. (2008). Fuel cells development and hydrogen production from renewable resources in Brazil. *International Journal of Hydrogen Energy*. 33. 4915-4935.

Li, Y., Chen, H., Ding, Y. (2010). Fundamentals and applications of cryogen as a thermal energy carrier: a critical assessment. *International Journal of Thermal Sciences*. 49. 941-949.

Li, Y., Chen, H., Zhang, X., Tan, C., Ding,Y. (2010). Renewable energy carriers: Hydrogen energy carriers: Hydrogen or liquid air/nitrogen?, *Applied Thermal Engineering*, 30, 1985-1990.

Lovins A. (2003) Twenty Hydrogen Myths. White Paper Document ID: E03-05, Publisher RMI. Available from: http://www.rmi.org/Knowledge-Center/Library/E03-05_TwentyHydrogenMyths

Meisen, P. (1996). Linking renewable energy resources around the world: a compelling global strategy. *Power Engineering Review*, IEEE 16, 14- 22.

Milciuviene, S., Milcius, D., Praneviciene, B. (2006). Towards hydrogen economy in Lithuania. *International Journal of Hydrogen Energy*. 31. 861 - 869.

Rabah, M.A., Eldighidy, S.M. (1989). Low cost hydrogen production from waste. *International Journal of Hydrogen Energy*.14. 221-229.

Ricci, M., Bellaby, P., Flynn, R. (2008). What do we know about public perceptions and acceptance of hydrogen? A critical review and new case study evidence, *International Journal of Hydrogen Energy*, 33, 5868-5880.

Ricci, M., Bellaby, P., Flynn, R. (2007). Stakeholders’ and publics’ perceptions of hydrogen energy technologies. In: Flynn R, Bellaby P, editors. *Risk and the public acceptance of new technologies*. Basingstoke: Palgrave Macmillan.

Ricci, M., Flynn, R., Bellaby, P. (2006). Public attitudes towards hydrogen energy: preliminary analysis of focus groups in London, Teesside and Wales. UKSHEC Social Science Working Paper 28. ISCPR, University of Salford. Available from, http://www.psi.org.uk/ukshec/pdf/28_attitudestohydrogen.pdf (02.06.2012).

Ricci, M., Flynn, R., Bellaby. P. (2006). Understanding the public acceptability of hydrogen energy – key findings from focus groups in Teesside, SW Wales and London (October– November 2006). UKSHEC Social Science Working Paper 33. University of Salford. Available from: http://www.psi.org.uk/ukshec/pdf/33_Ricci_etal_PubAcceptfinal%20report.pdf (28.05.2012).

Schmoyer, R.L., Truett, T., Cooper, C. (2004). Results of the 2004 knowledge and opinions surveys for the baseline knowledge assessment of the U.S. Department of Energy HydrogenProgram. ORNL/TM-2006/417. Oak Ridge National Laboratory. Available from: http://www1.eere.energy.gov/hydrogenandfuelcells/pdfs/survey_main_report.pdf (10.06.2012).

Solomon, B.D., Banerjee, D. (2006). A global survey of hydrogen energy research, development and policy. *Energy Policy*. 34. 781-792.

Turner, J., Sverdrup, G., Mann, M.K., Maness, P.C., Kroposki, B., Ghirardi, M., Evans, R.J., Blake, D. (2008). Renewable hydrogen production. *International journal of Energy Research*. 32. 379-407.

UNEP (United Nations Environment Program). (2006). *The hydrogen economy: a non-technical review*. Paris: UNEP Publications.

United States Department of Energy (2002) A National Vision of America's Transition to a Hydrogen Economy — to 2030 And Beyond, Based on the results of the National Hydrogen Vision Meeting Washington, DC November 15-16, 2001 (DOE, 2002). Available from: https://www1.eere.energy.gov/hydrogenandfuelcells/pdfs/vision_doc.pdf

A Vision of Our Future (Summary Report). *EC High Level Working Group on Hydrogen and Fuel Cells*, Brussels.

Waegel, A., Byrne, J., Tobin, D., Haney, B. (2006). Hydrogen highways: lessons on the energy technology–policy interface. *Bulletin of Science and Technology Society.* 26. 288 - 296.

Wietschel, M., Seydel, P. (2007). Economic impacts of hydrogen as an energy carrier in European countries. *International Journal of Hydrogen Energy,* 32, 3201-3211.

Wilk, R. (2007). Questionable assumptions about sustainable consumption. In: Reisch L, Ropke I, editors. The ecological economics of consumption. Cheltenham: Edward Elgar; 2004.perceptions of hydrogen energy technologies. In: Flynn R, Bellaby P, editors. Risk and the public acceptance of new technologies. Basingstoke: Palgrave Macmillan.

Wilsdon J, Willis R. (2004). See-through science: why public engagement needs to move upstream. London: Demos. Available from: http://www.demos.co.uk (29.05.2012).

Zachariah-Wolff, J.L., Hemmes, K. (2006). Public acceptance of hydrogen in the Netherlands: two surveys that demystify public views on a hydrogen economy. *Bulletin of Science, Technology & Society.* 32(4). 339–345.

Zhang, Y.-H.P. (2010). Renewable carbohydrates are a potential high density hydrogen carrier, *International Journal of Hydrogen Energy,* 35, 10334-10342.

Grid sampling with an application to a mixed-mode human survey

Seppo Laaksonen[1]

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

Abstract

Two types of strategies are used for sampling designing. One strategy is a standard stratified random sampling with regional strata, but the other uses special strata. These strata are based on 250mx250m grids so that all the grids are sorted by the income medians of the residents and two explicit strata are constituted. One of these grid strata consists of the grids with low income whereas the other of the grids with high income. These two types of samples are overlapping partially. However, there is need to use both samples in one framework. This leads to a non-trivial strategy for sampling and estimation.

Keywords: Strata, stratum overlapping, conditional inclusion probability

1 Introduction

The European Social Survey (ESS) is one of the most qualified surveys in Europe. Its sampling design varies from one country to the next, but we can still recognise the following basic features from these:

- Simple random sampling (srs) so that the study units (15+ aged residents) are explicitly available from a register.
- Random sampling with explicit strata, using often registers as well.
- Two stage cluster sampling so that the first stage units are small-area primary sampling units (psu's), whereas the two-stage units are directly as study units.
- Three stage cluster sampling so that the first-stage units are small-area primary sampling units (psu's), but the second-stage is needed to draw households or addresses before drawing the study units.

Srs naturally does not use stratification but the other three strategies often use, but not always. The main line in the ESS is that if stratification is used, it is explicit stratification and sample allocation is proportional, exactly or approximately. There are however many countries that use non-proportional allocation and even so that the anticipated response rates has an effect on the gross sample size. This has been made cautiously, for example so that the gross sampling fraction for large cities is higher than for rural areas. This thus, since the response rates seem to be low in large cities. In some cases, non-proportional allocation is made in order to obtain enough accurate results for certain explicit strata; the reason for this is national, it is not required by the ESS coordinating committee. See some information about the ESS sampling, Lynn et al 2007. The ESS website includes also useful information such as the sampling design principles (http://www.europeansocialsurvey.org/index.php?searchword=sampling& ordering=&searchphrase=all&Itemid=217&option=com_search).

We can consider the ESS as a standard survey in the sense that even though explicit strata are used, these are rather traditional such as administrative regions of a country. In this study, we go forward although we also use a very standard explicit stratification. On the other hand, our sample allocation is not proportional at all, but such that gives opportunity to get enough accurate estimates for specific strata. It should be noted that the use of anticipated response rates cannot be here used well, since our survey is rather unique and any a priori information does not exist. So, we hope that our 'intuition' for sample allocation was enough good from this point of view[1].

Administrative regions are important but people within these regions may be very different and the results obtained from these do not tell much about their attitudes or feelings, among others. Hence we try to go on to smaller areas and without administrative constraints. One strategy is to use Geographic Information System (GIS) so that small areas are the grids of 250 metres times 250 metres. People living within such small squares are expected to be as neighbours of each other, and hence their attitudes, feelings and opinions are maybe clustered to some extent. The whole data when being available in late 2012 give opportunity to analyse in very details urbanization vs ruralisation issues in the south Finland, in 16 municipalities totally, including Helsinki and its neighbour municipalities. This subject-matter analysis is forthcoming. This paper describes the sampling design strategies of the study.

The paper is organised so that we explain in Section 2 the target population and the sampling frame. Section 3 concentrates on the sampling design itself. It is good already to notice that we use the two different designs in fact. This leads to certain technical challenges that are solved in an interesting way in the next section. We present our solutions also empirically, using the gross sample data. Section 4 presents an interesting solution to calculate the inclusion probability. The final section discusses further steps that are possible to specify after the data from the respondents are available. The fieldwork has been conducted using such a mixed-mode design that gives for a potential respondent to participate either by web or by postal mail. This was considered to be best because it is inexpensive. In the forthcoming paper we also analyse the effects of this mixed-mode strategy that is rather new but becoming more common (see e.g. the ESS website: http://www.europeansocialsurvey.org/index.php?option=com_content&
view=article&id=67&Itemid=552).

2 Target population and sampling frame

The statistical units of the target population are 25-74 years old residents of 16 Finnish southern municipalities those mother tongue is either Finnish or Swedish. The information is based on the January 2012 population register. Our sampling frame has also constructed from this register.

From the regional point of view we have however two target populations, one being just those 16 municipalities. But the second is more complex and it is based on 250m x 250m grids of 14 out of these 16 municipalities. The reason for this is that two municipalities decide not to participate in this second study.

The first target population is divided into 19 explicit strata that are equal to the municipalities except that Helsinki consists of the three strata (most urbanised southern area, most urbanised northern area, suburb area). These are also administrative areas.

For the second regional target population, the income of the grids was used. The income concept is the taxable income from the 2010 taxation register. The median income of all the grids was computed and then the grids were sorted by this order, from the lowest median to the highest median. Consequently, two groups or strata were formed, the lowest quintile (called also 'poor') vs the highest quintile (called also 'rich'). This in-

[1] This study is initiated by Matti Kortteinen and Mari Vaattovaara from the University of Helsinki. My role has been and will be to help in methodological issues including sampling design that is the focus of this paper now.

formation was received from Statistics Finland who maintains the grid data base with population and taxation statistics data. Before determining the final strata, some robustness was made so that some initial grids were omitted. The basic reason was to protect people of too small grids. This was based on the confidentiality declaration of Statistics Finland.

When the set of grids was made robust, the two strata were ready to use. The first quintile thus constitutes one stratum and the fifth quintile the second, respectively. The map of Figure 1 shows how these two strata are spread around our municipalities. It is easy to see that ‘rich’ grids are concentrated on certain areas, and ‘poor’ grids on the other, respectively. However, any of them do not cover any whole municipality. There are empty areas from both types of grids, that is, their median income is somewhere in the middle (no poor, no rich) or the grids are ‘closed’ for confidentiality reasons.

Figure 1. Grids for ‘rich’ people vs. ‘poor’ people in the municipalities of the survey. The remaining grids are between those two ones

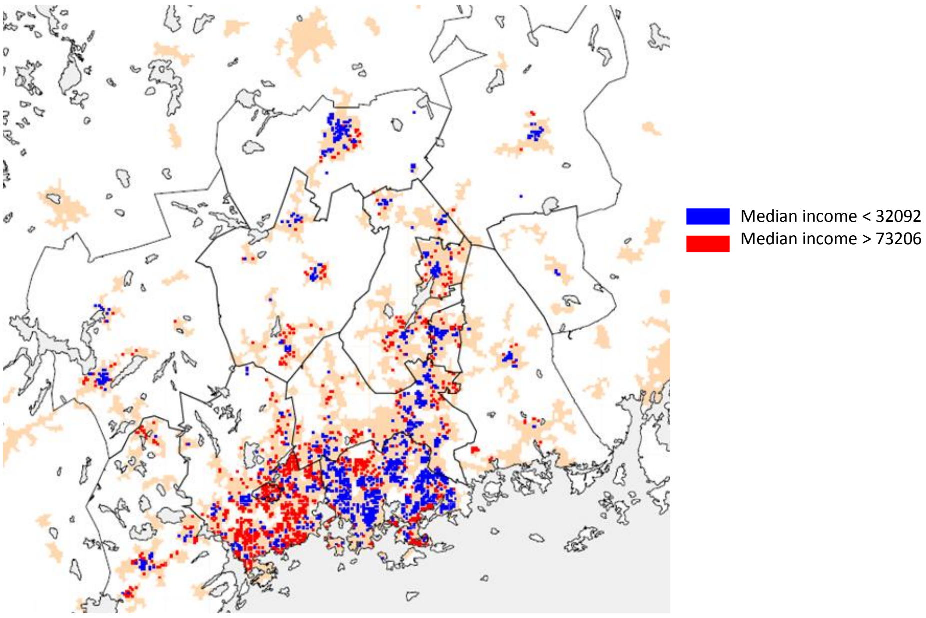


Table 1 shows what has been the ‘intuition’ of our research group. The grid-based stratum sizes were desired to be enough big in order to get enough accurate estimates. In the next section we come back to this issue and observe that the actual gross sample sizes are even higher due to sampling selection process used. The municipality based stratum sizes have been allocated much with a minimum principle that in our case means the 600 gross sample size, at minimum. Obviously this ensures that we will have enough respondents to estimate results reasonably well. If the response rate in such small municipalities would be for example 50%, we will get 300 respondents from this site of the data. This number is expected to increase from the grid site to some extent.

Table 1. Allocation of gross sample

Stratum	**Gross sample size**
Grids of 5th quintile income (High income grids, ‘Rich’)	6 000
Grids of 1th quintile income (Low income grids, ‘Poor’)	6 000
All income based strata	**12 000**
Espoo and Kauniainen	2 000
Helsinki, most urbanised southern area	1 000
Helsinki, most urbanised northern area	1 000
Helsinki, suburb	2 500
Hyvinkää	600
Järvenpää	600
Kauniainen	600
Kerava	600
Kirkkonummi	600
Lahti	1 000
Lohja	600
Mäntsälä	600
Nurmijärvi	600
Pornainen	600
Sipoo	600
Tuusula	600
Vantaa	1 500
Vihti	600
All municipality based strata	**15 000**
The whole gross sample	**27 000**

3 Sampling design

The sampling design for the both two parts of the survey is stratified random sampling. However, the design is not any standard such design, since these both samples are dependent. That is, the grid-part residents can be drawn to the sample also from the municipality part. Lahti and Lohja are the exceptions, their sampling design is exactly stratified random sampling.

The sample selection was performed by a sub-contractor who has access to the population register and to the grid information. The sub-contractor received the instructions to draw a sample but this could not be done so that all sampling principles were possible to take into account. The sample selection process was as follows.

First, the grid sample part was selected with the desired amount of respondents. This was done by addresses, so that one valid person from one address only was accepted. At the same time, this address selected was marked for the second round of the sampling selection that was concerned municipality samples. Thus, this second round was conditional to the first round, and hence it was not possible to draw the same person twice in the sample.

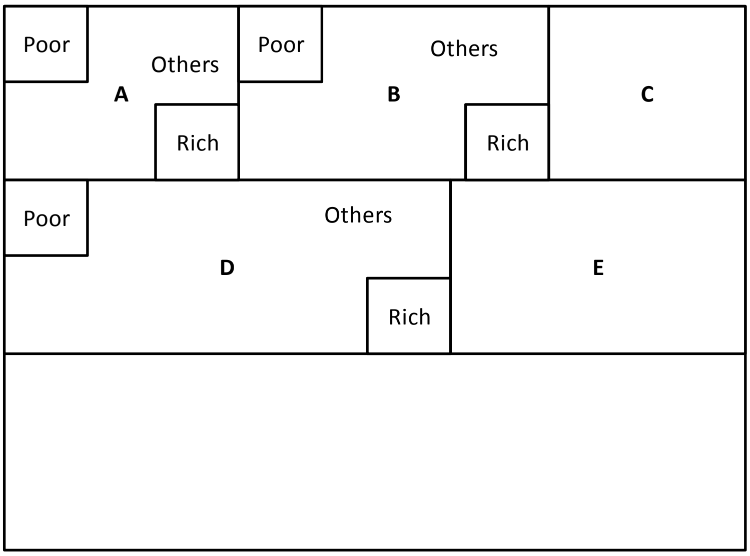
The inclusion probabilities are straightforwardly computable for Lahti and Lohja since any conditionality problem does not exist. They are as usually:

$$\pi_k = \frac{n_h}{N_h}$$

Here *h* is stratum (Lahti or Lohja), *n* is desired gross sample size and *N* = number of 15-74 years old residents, respectively.

The inclusion probabilities for the other municipalities and strata are more difficult to compute, since we have to know what is the probability for a selected person to be included in the sample? This probability was not available and hence we estimated it using the received gross sample that was possible after both samples were available. In order to illustrate the problem better, I use the following scheme:

Municipality Strata A, B, C, ..and Grid-based strata within each of them



First, we matched those two gross samples together so that each grid-sample person is identified to its municipality stratum. This was made by using postal zip codes of both data files. This identification worked well, although it was not definitely sure in advance. Now we were able to calculate the overlapping gross sample sizes and get good opportunities for estimating the inclusion probabilities. It is good to remind that sample selection is random within both grid strata, and it can be assumed that the distribution of the gross sample into municipalities corresponds approximately to the target population distribution. This assumption is in any way used in this study even not being perfectly true. It is clear that this uncertainty should be taken into account in variance estimation that is not included in this paper.

Table 2. Distribution of gross sample to strata. The group 'Others' in the above scheme is equal to municipality gross sample size.

	Poor grids	Rich grids	Municipality	Total	25-74 year Population
Helsinki, most urbanised southern area	110	46	1000	1156	27465
Helsinki, most urbanised northern area	1142	8	1000	2150	40206
Helsinki, suburb	2501	1324	2500	6325	147098
Espoo-Kauniainen	546	3127	2000	5673	131840
Hyvinkää	248	64	600	912	24944
Järvenpää	115	38	600	753	21717
Kerava	124	48	600	772	18874
Kirkkonummi	89	173	600	862	20065
Lahti	0	0	1000	1000	57059
Lohja	0	0	600	600	22613
Mäntsälä-Pornainen	49	22	600	671	13850
Nurmijärvi	85	120	600	805	21924
Sipoo	48	134	600	782	10269
Tuusula	118	201	600	919	20948
Vantaa	746	574	1500	2820	104930
Vihti	81	121	600	802	15923
All	6000	6000	15000	27000	699725

The inclusion probabilities are required to calculate separately to the three groups:

- for poor grids areas
- for rich grids areas
- for others who however can live either in poor grids, in rich grids or in intermediate poor/rich areas.

The sampling design for municipalities is independent of richness or poorness of their living grids, and hence the inclusion probabilities need to be calculated following this fact. For the analysis, it is of course possible to identify the respondents correctly by their grid. This information is also included in the data file.

In fact, we cannot calculate the inclusion probabilities straightforwardly. We had to 'estimate' them as explained below. I present them as the following formula:

$$\pi_k = \frac{n_{hc}}{\hat{N}_{hc}}$$

Here *c* is income level stratum (poor, rich, others). We have gross sample sizes for each strata as presented in Table 2, but we cannot know precisely population sizes for these overlapping strata. Hence we estimate them assuming that the gross sample size represents correctly to the corresponding population size. The formula for these statistics, e.g. for the stratum *h1* is as follows:

$$\hat{N}_{h1} = N_h \frac{n_{h1}}{(n_{h1} + n_{h2} + n_{h3})} .$$

4 Sampling design weights

When we have the inclusion probabilities, we can easily calculate the gross sample design weights:

$$w_k = \frac{1}{\pi_k}$$

Table 3 illustrates these design weights from our data. We see that the weights vary quite much that is due to the desired targets for the sample sizes. The variation for the grid part is smaller than for the municipality part.

Table 3. Some statistics of the gross sample design weights

Statistics	The whole sample	Grid part	Municipality part
Observations	27000	12000	15000
Mean	25.9	24.4	27.1
Total	699725	292615	407110
Minimum	13.1	13.1	13.1
Maximum	57.1	37.2	57.1
CV (%)	31.7	20.5	36.4

Note: The overlapping is useful thus for the our big point, to compare people's attitudes, living conditions etc within different types of very small areas, such as 250m x 250 grids. I already mentioned that the gross sample size (and net sample size consequently) will be increased from the initial targets due to the overlapping. When identifying people of the municipality sample into poor vs rich grids, our gross sample size was increased essentially, from 6000 to 9572 in poor grids, but only from 6000 to 6992 in rich grids. We can thus observe that a random selection provides relatively much more people from poor grids than from rich grids.

5 Concluding remarks and future

This is a new and obviously innovative approach to survey sampling, especially for stratification. The GIS data are used for many purposes but not so much for sampling and estimating. At least, I have not seen the approach like this in literature.

Our respondent data are soon becoming to be available. This gives opportunity to create the sampling weights for the respondents. Such initial or base weights are easy to compute, that is, just to change gross sample sizes n to the corresponding net sample sizes, let say r. This weighting is only the start for constructing good sampling weights. These require also to analyse non-response and to adjust for it. My plan is to use the response propensity modelling first and then to calibrate the sums of the resulted weights into the sums of the gross sample weights. This will be done at each stratum level so that overlapping strata are covered too (e.g. Laaksonen 2007).

The response propensity modeling is more advantageous if good auxiliary variables are available. Our pattern is not perfect, thanks for the problem that we are outside Statistics Finland who has more such variables easily available. We have not obtained for example education that is too hard to get for outsiders, but we have many population register variables fortunately, such as age, gender, mother tongue, dwelling unit structure, previous living area, house type and house size. Our group is also going to ask basic information from the taxation register and the employment register.

References

Laaksonen, S. (2007). Weighting for Two-Phase Surveyed Data. *Survey Methodology*, December Vol. 33, No. 2, pp. 121-130, Statistics Canada.

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.

Challenges and possibilities of usage MICS in Belarus

Anna Larchenko[1]

[1]Belarus State Economic University, e-mail: annalarchenko@gmail.com

Abstract

The history and methodology of MICS in Belarus is considered. Major changes in MICS 2012 are investigated. The main shortcomings of MICS in Belarus are revealed.

Keywords: Multiple Indicator Cluster Survey, sample unit, households sample survey, sample size.

1 Introduction

MICS covers many countries and it's carried out since the mid-1990, under the auspices of UNICEF.

MICS 1 was developed in response to the World Summit for Children to measure progress towards an internationally agreed set of mid-decade goals. The first round of MICS was conducted around 1995 in more than 60 countries.

MICS 2 was conducted in 2000, and it included about 65 surveys.

MICS 3 covered about 50 countries including Belarus.

Results from ***MICS 4*** surveys, carried out in 2009-2011, will allow countries to better monitor progress toward national goals and global commitments, including twenty of the Millennium Development Goals (MDGs) as the target year 2015 approaches.

2 MICS 3 and MICS 4 in Belarus

MICS 3 in Belarus. The object of the study included children aged under five years old and women aged 15-49 years in the total households sample (7000 households (HH), including those with the children under 5 years old – 2870).

The purpose of the survey was obtaining statistical data to assess the status of women in reproductive age and children under five years old.

In the construction of the sample the method of random sampling without replacement was used. The sample was based on the lists of HH addresses and the information of medical institutions located in the surveyed settlements, and providing child care.

Survey instruments consisted of three *questionnaires*:

- "Household Questionnaire";
- "Individual Questionnaire For Women Aged 15-49 Years";
- "Questionnaire For Children Under Five Years Old".

Each of them consisted of several modules (sets of questions on a specific topic). Thus, "Household Questionnaire" included the following modules: information about the household, household inventory, education, water and sanitation, household characteristics, child labour; "Individual Questionnaire For Women Aged 15-49 Years" included topics: Information about the woman, infant mortality, maternal and neonatal health, marital status, contraceptive use, HIV / AIDS; "Questionnaire for Children Under Five Years Old" consisted of: early training, breastfeeding, disease treatment and child care, immunization, children's anthropometric data.

MICS 4 in Belarus. In March - June 2012 another survey was held in Belarus to assess the status of children and women.

For this survey four *questionnaires* have been designed:

- "Household Questionnaire";
- "Individual Questionnaire For Women Aged 15-49 Years";
- "Individual Questionnaire For Men Aged 15-59 Years";
- "Questionnaire For Children Under Five Years Old".

The module "Reproductive Health" has been included in the questionnaire for women for the first time. However, the range of questions on the given subjects is very limited. Into the questionnaire for women the other modules have been also added: access to media and information and communication technologies, desirability of the last born child, monitoring during the postpartum period, symptoms of diseases, relation to domestic violence, sexual behavior, tobacco and alcohol use, life satisfaction; into the questionnaire for households the following modules has been added: discipline of children, salt iodization.

The purpose of the survey is obtaining at the national level, as well as by regions, the data to assess the level of living conditions, health status of women in reproductive age, children under 5 years and men 15-59 years of age.

Sample unit is a household.

For a total population a set of private HH of Belarus is accepted. Collective households (1.1% of the total population), students living in residence (1.7%) and homeless (less than 0.1%) are excluded.

When determining the *sample size* the following formula is used:

$$n = \frac{4(r)(1-r)(f)(1.1)}{(0.12r)^2(p)(n_h)}, \tag{1}$$

where n is a required sample size; 4 is the coefficient, providing 95 percent confidence level; r – predicted or expected prevalence (coverage rate) of the indicator; 1.1 – the coefficient that is required to increase the sample size by 10% for non-response compensation; *f* – deff; 0.12*r* – the margin of error acceptable at the 95-percent confidence level, defined as 12% of *r* (a relative sampling error for the *r*); *p* – proportion in the total population, which is based on the parameter *r*; n_h is the average household size.

When calculating it was assumed that:

- *r* (hypothetical prevalence of any key indicator) is 50%;

- *f* is 1.5;

- *p* (the proportion of children aged 0-4 years in the total population) is 5.2%;

- n_h (average household size) is 2.43.

A result of calculations has shown the required sample size for each territory was 3627 HHs. Thus, the number of HHs in Belarus as a whole is equal to 25,389 HH (i.e., 3627×7 = 25,389).

However, considering financial expenditures and the limited time of the survey, it was decided to use the existing in the country sample set that is used for the survey of households' living standards (6000 HH).

However, due to the low average size of household (2.4 by the Census of 2009), a small weight of children under 5 years in the population (5.2%) and the age of 2 years (2, 2%) a limited number of children under five years are represented in the sample. In this regard, an additional subsample of households with children aged 0-4 years was formed.

To calculate *the number of HHs with children under 5 years* the following formula was used:

$$n = \frac{4(r)(1-r)(f)(1.05)}{(0.12r)^2(l)}, \qquad (2)$$

where *n* is a required quantity of children in the sample; 4 is the coefficient, providing 95 percent confidence level; *r* – expected prevalence rate; 1.05 – the coefficient that is required to increase the sample size by 5% for non-response compensation; *f* – deff; 0.12*r* – the margin of error acceptable at the 95-percent confidence level, defined as 12% of *r* (a relative sampling error for the *r*); *l* is the size of target group of children on the average per household with children under the age of five years.

When calculating it was assumed that:

- *r* is 50%;

- *f* is 1.5;

- *l* is 1.13 (based on estimates obtained from the survey of households' living standards).

Three-stage territorial probability stratified sampling has been used; equiprobable selection method has been applied.

To ensure uniform distribution of the sample set, selection has been made separately for the Brest, Vitebsk, Gomel, Grodno and Minsk regions and Minsk-city.

The selection of units has been made in three stages:

- at the first stage the primary sampling units included administrative-territorial items: cities, towns, village councils. To form a representative sample set and to ensure a relative homogeneity of the groups the stratification of the total population has been carried out;

- at the second stage the following units have been taken: in cities and towns – Census plots, in rural areas – a set of settlements within the rural councils. The number of selected clusters in each region is given in Table 1.

Table 1 – The number of clusters selected by regions and Belarus as a whole

	Total	Including		
		«large» cities	«small» cities	rural councils
Republic of Belarus	343	204	86	53
Regions:				
Brest	49	24	15	10
Vitebsk	49	30	11	8
Gomel	56	36	12	8
Grodno	38	15	15	8
Minsk-city	63	63	-	-
Minsk	46	12	21	13
Mogilev	42	24	12	6

- at the third stage the sampling unit was household.

In such a way the survey has covered more than 7800 HH, 2710 of them – households with children under the age of 5 years.

The publication of a preliminary report on the survey is planned for September-October 2012, a final – in December 2012 – January 2013.

3 Concluding remarks

The use of three-stage territorial cluster sampling provides rather reliable information across more indicators of MICS, conducted in Belarus.

However, despite the wide range of indicators derived from the MICS, there are several shortcomings of the survey:

- it is carried out not regularly (every five-seven years);

- the module "Reproductive Health" is introduced in the survey for the first time and is only available for women;

- the wealth of information is insufficient for a comprehensive evaluation of reproductive population health.

References

Multiple Indicator Cluster Survey of the living conditions for children and women aged 15-49 in 2005: final report. – Minsk: Ministry of Statistics and Analysis of Belarus; Research Institute of Statistics; Children's Fund of the United Nations (UNICEF). – 2007.

On organizing of household sample survey in the Republic of Belarus [Electronic resource] / National Statistical Committee of the Republic of Belarus. http://belstat.gov.by/homep/ru/households/1.php.

Multiple Indicator Cluster Survey (MICS) [Electronic resource] / UNICEF. http://www.unicef.org/statistics/index_24302.html.

Estimation Under Restrictions Built Upon Biased Initial Estimators

Natalja Lepik[1]

[1]University of Tartu, e-mail: natalja.lepik@ut.ee

Abstract

The users of official statistics often require that sample-based estimates satisfy certain restrictions. In the domain's case it is required that the estimates of domain totals sum up to the population total or to its estimate. The general restriction estimator (GR) proposed by Knottnerus (2003) is described in this paper, which uses an unbiased initial estimators for its construction. Also three new estimators that satisfies the linear restriction are proposed and compared. We allow the initial estimators for them to be biased.

Keywords: Survey sampling, restriction estimator

1 Introduction

Nowadays, demand on accurate statistics of population sub-groups or domains increases. This statistics can be obtained from surveys, or, sometimes, aggregated from registers. It may happen that even if the register contains variables under interest, it does not contain identifies of the domains under our particular interest. As follows, these domain totals can not be produced from that register, they need to be estimated from a survey. The survey has to collect information on the same study variable but together with domain identifiers. As a result, the consistency problem occurs, the domain estimates from the survey do not sum up to the totals available from the registers. Analogical problem arises in the multi-survey situation, where some study variables are common in two or more surveys. Domain estimates from one survey do not sum up to the estimates of larger domains (or population totals) from another survey. Yet, there is one more situation where the consistency problem occurs. Domains themselves and the population total may be estimated by conceptually different estimators in the same survey. As a result, the domain totals do not sum up to the population total, or to the relevant larger domains.

The described inconsistency is annoying from the statistics users viewpoint. Statisticians know that the relationships between population parameters do not necessarily hold for the estimates in a sample. They also know that any auxiliary information incorporated into estimators may increase precision of these estimators. In our situation known relationships between population parameters is a kind of the auxiliary information. Involving this information into estimation process presumably improves estimates. Our goal is to define consistent domain estimators that are more accurate than the initial inconsistent domain estimators.

The problem is not new, consistency of estimators has been considered for some time. For example, if consistency is required between two surveys or between a survey and a register, some authors (Zieschang 1990, Renssen and Nieuwenbroek 1997, Traat and Särndal 2009, Dever and Valliant 2010) have proposed classical calibration approach as a solution. In this approach, the common variables are considered as additional auxiliary variables, and consistency requirement is presented in terms of calibration constraints. Other authors (Kroese and Renssen 1999, Knottnerus and Van Duin 2006) use different calibration approach for this situation, called repeated weighting. They re-calibrate the initially calibrated estimators to satisfy the consistency constraints with outside information.

Yet another approach is proposed by Knottnerus (2003). His estimator is based on the unbiased initial estimators and is unbiased itself. The advantage of the GR estimator is the variance minimizing property in a class of linear estimators. Sõstra (2007) has developed the GR estimator for estimating domain totals under summation restriction. Optimality property of the domain GR estimator is studied in Sõstra and Traat (2009). In all these works, the unbiased or asymptotically unbiased initial estimators are assumed.

It is well known that there are many useful estimators that are biased. For example, the model-based small area estimators are design-biased. The synthetic estimator can be biased on the domain level. Even the widely used GREG estimator is only asymptotically unbiased. In this paper we will allow the vector of initial estimators $\hat{\boldsymbol{\theta}}$ to be biased, and will construct three new restriction estimators, based on the biased initial estimators.

2 Estimation under restriction

Let finite population U be divided into D non-overlapping domains U_d, $d \in \mathcal{D} = \{1, 2, ..., D\}$. We are interested in some domain parameters, for example domain totals, $t^d = \sum_{i \in U_d} y_i$ with y_i being the value of study variable for object i. It is natural that domain totals sum up to the population total, $\sum_{d=1}^{D} t^d = t = \sum_{i \in U} y_i$.

2.1 Knottnerus' approach

In general case, we denote the parameter vector under study by $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)'$, it satisfies linear restrictions:

$$\mathbf{R}\boldsymbol{\theta} = \boldsymbol{c}, \tag{1}$$

where $\mathbf{R}$ is an $r \times k$ matrix of rank r and $\boldsymbol{c}$ is the r-dimensional vector of known constants.

In a case of domain totals, where $\sum_{d=1}^{D} t^d = t$,

$$\mathbf{R} = (1, 1, ..., 1, -1)_{1 \times (D+1)},\ \boldsymbol{\theta} = (t_y^1, t_y^2, ..., t_y^D, t_y)' \text{ and } \boldsymbol{c} = 0,$$

or alternatively,

$$\mathbf{R} = (1, 1, ..., 1)_{1 \times D},\ \boldsymbol{\theta} = (t_y^1, t_y^2, ..., t_y^D)' \text{ and } \boldsymbol{c} = t_y.$$

Let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_k)$ be the vector of estimators of $\boldsymbol{\theta}$ that do not necessarily satisfy the linear restriction (1), i.e. $\mathbf{R}\hat{\boldsymbol{\theta}} \neq \boldsymbol{c}$, in general. Knottnerus (2003, p. 328-329) proposes the following restriction estimator to solve this problem.

Assume that $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_k)'$ is unbiased for the parameter vector $\boldsymbol{\theta}$ with the variance $\mathbf{V}$, such that $\mathbf{RVR}'$ can be inverted. Then the general restriction estimator $\hat{\boldsymbol{\theta}}_{GR}$ that satisfies restrictions (1) for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{GR}$, and the variance $\mathbf{V}_{GR}$ of this estimator are:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{GR} &= \hat{\boldsymbol{\theta}} + \mathbf{K}(\boldsymbol{c} - \mathbf{R}\hat{\boldsymbol{\theta}}), & (2) \\ \mathbf{V}_{GR} &= \mathbb{C}\text{ov}(\hat{\boldsymbol{\theta}}_{GR}) = (\mathbb{I} - \mathbf{KR})\mathbf{V}, & (3) \end{aligned}$$

where $\mathbb{I}$ is the $k \times k$ identity matrix and

$$\mathbf{K} = \mathbf{VR}'(\mathbf{RVR}')^{-1}. \tag{4}$$

Since $\mathbf{RK}$ is the identity matrix, it is easy to check that $\hat{\boldsymbol{\theta}}_{GR}$ satisfies restrictions (1):

$$\mathbf{R}\hat{\boldsymbol{\theta}}_{GR} = \mathbf{R}\hat{\boldsymbol{\theta}} + \mathbf{RK}(\boldsymbol{c} - \mathbf{R}\hat{\boldsymbol{\theta}}) = \boldsymbol{c}.$$

Knottnerus (2003, p. 332) shows that $\hat{\boldsymbol{\theta}}_{GR}$ is optimal in a class of estimators that are linear in $\hat{\boldsymbol{\theta}}$ and satisfy restrictions (1). In this class, $\hat{\boldsymbol{\theta}}_{GR}$ has minimum variance (in Löwner ordering). For example, other estimators in this class can be received by replacing $\mathbf{V}$ in the expression of $\mathbf{K}$ by any arbitrary $k \times k$ matrix $\mathbf{V}^*$, such that $\mathbf{RV}^*\mathbf{R}$ can be inverted. But the resulting estimators have bigger variance than

$\hat{\boldsymbol{\theta}}_{GR}$. In Sõstra (2007, p. 45) it is also shown that $\hat{\boldsymbol{\theta}}_{GR}$ is never less efficient than the initial estimator $\hat{\boldsymbol{\theta}}$, $\mathbf{V}_{GR} \leq \mathbf{V}$ in the sense of Löwner ordering.

Without loss of generality, we further consider linear restrictions in the form

$$\mathbf{R}\boldsymbol{\theta} = \mathbf{0}. \tag{5}$$

With $\boldsymbol{c} = \mathbf{0}$, the Knottnerus' GR estimator simplifies to the form

$$\hat{\boldsymbol{\theta}}_{GR} = (\mathbb{I} - \mathbf{KR})\hat{\boldsymbol{\theta}}. \tag{6}$$

For biased estimators the accuracy of the estimator is ordinarily measured by its mean square error. The GR-estimator (6) with biased initial estimator $\hat{\boldsymbol{\theta}}$ is not optimal any more for $\boldsymbol{\theta}$ in the sense of MSE. Although it still satisfies restrictions (5), it may have bigger mean square error than that of the initial estimator. For further details see Lepik (2011, p. 36).

In the following section we allow initial estimator to be biased, and we define three different restriction estimators for this case.

2.2 Restriction estimators handling bias

Assume that estimator $\hat{\boldsymbol{\theta}}$ is biased for $\boldsymbol{\theta}$,

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta} + \boldsymbol{b}, \tag{7}$$

where $\boldsymbol{b}$ is a vector of biases.

The first restriction estimator with biased initial estimators is defined in the following proposition.

Proposition 1. *The estimator*

$$\hat{\boldsymbol{\theta}}_{GR1} = (\mathbb{I} - \boldsymbol{KR})(\hat{\boldsymbol{\theta}} - \boldsymbol{b}), \tag{8}$$

with $\boldsymbol{K} = \boldsymbol{VR}'(\boldsymbol{RVR}')^{-1}$ *is unbiased for* $\boldsymbol{\theta}$. *Its variance is*

$$\mathbb{C}ov(\hat{\boldsymbol{\theta}}_{GR1}) = (\mathbb{I} - \boldsymbol{KR})\,\boldsymbol{V}, \tag{9}$$

and it is the optimal estimator among all linear estimators in $(\hat{\boldsymbol{\theta}} - \boldsymbol{b})$ *that satisfy restriction (5).*

For the proofs of this result and the following propositions see Lepik (2011, pp. 38-42).

Similarly to Knottnerus GR estimator our $\hat{\boldsymbol{\theta}}_{GR1}$ requires quantities that are usually unknown in practise, here the bias $\boldsymbol{b}$ and the variance $\mathbf{V}$. If $\mathbf{V}$ and $\boldsymbol{b}$ are replaced with consistent estimators, $\hat{\boldsymbol{\theta}}_{GR1}$ is consistent itself.

Below we define an estimator that is free of the knowledge of $\boldsymbol{b}$, satisfies restrictions and is more accurate than the initial estimator $\hat{\boldsymbol{\theta}}$, in $\mathbb{M}$SE terms.

Proposition 2. *The estimator, satisfying restrictions (5), but based on the mean square error* $\boldsymbol{M}$ *of the initial estimator* $\hat{\boldsymbol{\theta}}$, *is*

$$\hat{\boldsymbol{\theta}}_{GR2} = (\mathbb{I} - \boldsymbol{K}^{*}\boldsymbol{R})\hat{\boldsymbol{\theta}}, \tag{10}$$

where $\boldsymbol{K}^{*} = \boldsymbol{MR}'(\boldsymbol{RMR}')^{-1}$. *The bias of the* $\hat{\boldsymbol{\theta}}_{GR2}$ *is*

$$\boldsymbol{b}(\hat{\boldsymbol{\theta}}_{GR2}) = (\mathbb{I} - \boldsymbol{K}^{*}\boldsymbol{R})\boldsymbol{b}, \tag{11}$$

and the mean square error matrix is

$$\mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR2}) = (\mathbb{I} - \boldsymbol{K}^{*}\boldsymbol{R})\boldsymbol{M}. \tag{12}$$

Furthermore,

$$\mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR2}) \leq \boldsymbol{M} \tag{13}$$

in the sense of Löwner ordering.

The third estimator with its properties is proposed in the following proposition.

Proposition 3. *The restriction estimator*

$$\hat{\boldsymbol{\theta}}_{GR3} = (\mathbb{I} - \boldsymbol{K}^{*}\boldsymbol{R})(\hat{\boldsymbol{\theta}} - \boldsymbol{b}) \tag{14}$$

with $\boldsymbol{K}^{*} = \boldsymbol{MR}'(\boldsymbol{RMR}')^{-1}$ *satisfies restrictions (5) and is unbiased for* $\hat{\boldsymbol{\theta}}$*. It's MSE is the covariance of the estimator and is equal to*

$$\mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR3}) = (\mathbb{I} - \boldsymbol{K}^{*}\boldsymbol{R})\,\boldsymbol{V}(\mathbb{I} - \boldsymbol{K}^{*}\boldsymbol{R})'. \tag{15}$$

Furthermore,

$$\mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR3}) \leq \boldsymbol{M}. \tag{16}$$

It easy to ensure that GR estimator (6) is the particular case of the estimators $\hat{\boldsymbol{\theta}}_{GR1}$, $\hat{\boldsymbol{\theta}}_{GR2}$ and $\hat{\boldsymbol{\theta}}_{GR3}$, if the vector of initial estimators has the zero bias, $\boldsymbol{b} = \boldsymbol{0}$. These estimators have higher accuracy than the initial estimator $\hat{\boldsymbol{\theta}}$ in a term of MSE. The next result compares the accuracy of all four estimators.

Proposition 4. *The mean square error matrices of the restriction estimators* $\hat{\boldsymbol{\theta}}_{GR1}$*,* $\hat{\boldsymbol{\theta}}_{GR2}$*,* $\hat{\boldsymbol{\theta}}_{GR3}$ *and the initial estimator* $\hat{\boldsymbol{\theta}}$ *can be ordered (in the sense of Löwner ordering) as following:*

$$\mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR1}) \leq \mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR3}) \leq \mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR2}) \leq \mathbb{M}SE(\hat{\boldsymbol{\theta}}). \tag{17}$$

2.3 Some thoughts for the future research

Estimators GR1 and GR3 requires the knowledge of the bias $\boldsymbol{b}$. In practise it is usually not known, sometimes can be estimated. The behavior of the estimators $\hat{\hat{\boldsymbol{\theta}}}_{GR1} = (\mathbb{I} - \mathbf{KR})(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{b}})$ and $\hat{\hat{\boldsymbol{\theta}}}_{GR3} = (\mathbb{I} - \mathbf{K}^{*}\mathbf{R})(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{b}})$ is not studied yet.

Analogical situation is with the quantities $\mathbf{V}$ and $\mathbf{M}$. If to replace these quantities with their unbiased estimates, then the ordering of the MSEs of GR1, GR2 and GR3 is not necessarily hold.

References

Dever, J.A., Valliant, R.L. (2010) A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology*, 36(1), pp. 45-56.

Knottnerus, P. (2003) *Sample Survey Theory. Some Pythagorean Perspectives.* Wiley, New York

Knottnerus, P., van Duin, C. (2006). Variances in Repeated Weighting With an Application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22, pp. 565-584.

Kroese, A.H., Renssen, R.H. (1999). Weighting and Imputation at Statistics Netherland. *Proceedings of the IASS Conference on Small Area Estimation*, Riga, 109-120.

Lepik, N. (2011) *Estimation of domains under restrictions built upon generalized regression and synthetic estimators. Doctoral Dissertation.* Tartu

Renssen, R.H., Nieuwenbroek, N.J. (1997), Aligning Estimates for Common Variables in two or More Sample Surveys, *Journal of the American Statistical Association*, 92, 368-374.

Sõstra, K. (2007) *Restriction estimation for domains. Doctoral Dissertation.* Tartu

Sõstra, K., Traat, I. (2009) Optimal domain estimation under summation restriction. *Journal of Statistical Planning and Inference* vol. 139, pp. 3928-3941

Traat, I., Särndal, C.E. (2009). Domain Estimators Calibrated on Information from Other Surveys. *Research Report* No. 2009-1, Vol. 15, Department of Mathematics and Mathematical Statistics, Umea University, Sweden.

Zieschang, K.D. (1990), Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.

The Simulation Study of Survey Cost and Precision

Mārtiņš Liberts[1]

[1]University of Latvia, e-mail: martins.liberts@gmail.com

Abstract

Cost efficiency is a desirable property for sample surveys done in practice. It is a common task for a statistician to find the balance between precision and cost during the planning stage of a survey. For example, cluster sampling can be preferable choice regarding cost efficiency because the reduction of cost can dominate the loss in precision. Assessment of survey design regarding cost efficiency can be complex task. The approach presented in the talk is to use the methodology of simulation experiments as a tool for the cost efficiency analysis. Artificial population data is used in simulation experiments. The artificial population is created using the data from the Population Register of Latvia and the data from Latvian Labour Force Survey (LFS). The analysis of sampling design used for Latvian LFS will be presented. Two stage sampling design is used for the Latvian LFS where census counting areas are primary sampling units and dwellings are secondary sampling units. The design will be compared with other traditional sampling designs regarding cost efficiency.

Keywords: Survey sampling, simulation, cost, precision

1 Introduction

The idea of the study comes from purely practical necessity. National statistical institutes (NSI) usually are the main providers of the official statistics. The customers of the official statistics are society or tax payers in other words. Cost efficiency is one of the very desirable property for the government spendings. Somebody can ask a question – is the survey organised by NSI is cost efficient?

The aim the of the study is to develop a practical tool to compare different sampling strategies regarding a cost efficiency. The work is based on the Labour Force Survey.

2 The target population

The target population of the Labour Force Survey (LFS) usually is defined as all residents permanently living in private households (age group of the working age – 15-74 is the main domain of the interest). The target population is constantly changing over time. The target population is observed on weekly bases by the methodology of LFS (European Communities, 2003). Questioning of all residents every week would be required if LFS would be done as a census (full survey of whole target population). The example of the LFS target population is given by the table 1.

The population can be represented as a table with rows representing individuals and columns representing weeks. There are N individuals labelled with labels $1, 2, \dots, N$. The populations in the table 1 refers to W weeks. Weeks are labelled with labels $1, 2, \dots, W$.

There is an assumption of fixed set of individuals during the period of W weeks. It means there is not any "birth" or "death" during the time period under consideration. This assumption does not hold in practice.

Table 1: The target population of LFS

i	w=1	w=2	w=3	w=4	w=5	$\cdots$	w=W
1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$\cdots$	$y_{1,W}$
2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$\cdots$	$y_{2,W}$
3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	$y_{3,5}$	$\cdots$	$y_{3,W}$
4	$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$	$y_{4,5}$	$\cdots$	$y_{4,W}$
5	$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$	$y_{5,5}$	$\cdots$	$y_{5,W}$
6	$y_{6,1}$	$y_{6,2}$	$y_{6,3}$	$y_{6,4}$	$y_{6,5}$	$\cdots$	$y_{6,W}$
$\cdots$							
N	$y_{N,1}$	$y_{N,2}$	$y_{N,3}$	$y_{N,4}$	$y_{N,5}$	$\cdots$	$y_{N,W}$

3 Parameters of interest

Two population parameters are considered – total and ratio of two totals.

3.1 Total

Weekly total for the week w is defined by the equation 1.

$$Y_w = \sum_{i=1}^{N} y_{i,w} \tag{1}$$

Quarterly total for the quarter q is defined by the equation 2. There is an assumption – all quarters consist of 13 weeks. There are some quarters with 14 weeks in real calendar.

$$Y_q = \frac{1}{13}\sum_{w=j}^{j+12} Y_w = \frac{1}{13}\sum_{w=j}^{j+12}\sum_{i=1}^{N} y_{i,w} \tag{2}$$

Yearly total for the year y is defined by the equation 3. There is an assumption – all years consist of 4 quarters or 52 weeks. There are some years with 53 weeks in real calendar.

$$Y_y = \frac{1}{4}\sum_{q=k}^{k+3} Y_q = \frac{1}{52}\sum_{w=j}^{j+51} Y_w = \frac{1}{52}\sum_{w=j}^{j+51}\sum_{i=1}^{N} y_{i,w} \tag{3}$$

3.2 Ratio of two totals

Weekly ratio of two totals for the week w is defined by the equation 4.

$$R_w = \frac{Y_w}{Z_w} = \frac{\sum_{i=1}^{N} y_{i,w}}{\sum_{i=1}^{N} z_{i,w}} \tag{4}$$

Quarterly ratio of two totals for the quarter q is defined by the equation 5.

$$R_q = \frac{Y_q}{Z_q} = \frac{\sum_{w=j}^{j+12} Y_w}{\sum_{w=j}^{j+12} Z_w} \tag{5}$$

Yearly ratio of two totals for the year y is defined by the equation 6.

$$R_y = \frac{Y_y}{Z_y} = \frac{\sum_{q=k}^{k+3} Y_q}{\sum_{q=k}^{k+3} Z_q} = \frac{\sum_{w=j}^{j+51} Y_w}{\sum_{w=j}^{j+51} Z_w} \tag{6}$$

4 Design efficiency

4.1 The balance of variance and cost

Assume an arbitrary population parameter θ. Assume there is a probability sample s drawn by known sampling design $p(s)$. θ can be estimated using an estimator $\hat{\theta}_p$. The variance of $\hat{\theta}_p$ is denoted by $V\left(\hat{\theta}_p\right)$.

Assume a cost associated to a sample s. This is a cost what survey organiser has to spend to carry out the survey with sample s. A cost can be expressed in money, time or other quantity. Assume there is a cost function $c(s)$. The cost of sample s can be computed by the cost function $c_s = c(s)$. c_s is a random because s is a random sample. The expectation of c_s under sampling design $p(s)$ is notated as $E(c_s) = C_p$.

Usual desire is to minimise $V\left(\hat{\theta}_p\right)$ and C_p. Unfortunately these are conflicting tasks. You have to increase cost to reduce the variance and variance goes up when cost is reduced. The usual task of statistician is to construct sampling design $p(s)$ so that C_s and $V\left(\hat{\theta}_p\right)$ would be in "balance".

4.2 Design effect

There is a need for a measure of design efficiency. Assume two sampling designs:

- Simple random sampling – *srs*
- Alternative sampling design – $p(s)$

The classical design effect is a ratio of variances under condition of equal sample sizes defined by the equation 7.

$$deff\left(p, \hat{\theta}, n\right) = \frac{V\left(\hat{\theta}_p | E(n_p) = n\right)}{V\left(\hat{\theta}_{srs} | n_{srs} = n\right)} \tag{7}$$

$\hat{\theta}_p$ denotes π estimator under sampling design p(s), $\hat{\theta}_{srs}$ denotes π estimator under simple random sampling.

Alternative design effect can be introduced by the equation 8. It is defined as a ratio of variances under condition of equal expected costs.

$$deff^{\star}\left(p, \hat{\theta}, \gamma\right) = \frac{V\left(\hat{\theta}_p | C_p = \gamma\right)}{V\left(\hat{\theta}_{srs} | C_{srs} = \gamma\right)} \tag{8}$$

The design effect defined by (8) could be used as a measure of design efficiency. Assume two sampling designs – $p(s)$ and $q(s)$.

Definition 1 *The sampling design $p(s)$ is more efficient then the sampling design $q(s)$ for estimation of θ with survey budget γ if $deff^{\star}\left(p, \hat{\theta}, \gamma\right) < deff^{\star}\left(q, \hat{\theta}, \gamma\right)$.*

The definition 1 is equivalent to the definition 2:

Definition 2 *The sampling design $p(s)$ is more efficient then the sampling design $q(s)$ for estimation of θ with survey budget γ if $V\left(\hat{\theta}_A, C_A = \gamma\right) < V\left(\hat{\theta}_B, C_B = \gamma\right)$.*

5 Simulation

Design efficiency can be measured with help of simulation experiments. Artificial population data are necessary to carry out the simulation experiments. The artificial population data are created from the data of the Latvian Population Register and the survey data of Latvian LFS.

5.1 Sampling designs

The task of this research is to measure the design efficiency for three sampling designs used to select a quarterly sample – sample for 13 weeks.

There are two questionnaires for LFS – household questionnaire and individual questionnaire.

5.1.1 SRS of individuals

The SRS of individuals is selected. Sample of individuals is allocated randomly and evenly over 13 weeks.

The sample of dwellings is constructed from the sample of individuals. A dwelling is sampled if at least one dwelling member is sampled. A household questionnaire is filled for each sampled dwelling. An individuals questionnaire is filled for each sampled individual.

5.1.2 SRS of dwellings

The SRS of dwellings is selected. Sample of dwellings is allocated randomly and evenly over 13 weeks.

A household questionnaire is filled for each sampled dwelling. Individuals questionnaires are filled for all individuals from a sampled dwelling.

5.1.3 Two stage sampling design

This is the sampling design used in practice for Latvian LFS (Liberts, 2010).

A household questionnaire is filled for each sampled dwelling. Individuals questionnaires are filled for all individuals from a sampled dwelling.

5.2 Cost function

The cost is expressed as time necessary for field interviewers to carry out the survey in the simulation. There two components:

- Time for travelling $t_1(s) = \frac{\sum_{g=1}^{G} d_g}{\bar{v}}$ where G is a number of interviewers, d_g is a distance done by interviewer g to carry out the survey, $\bar{v}$ – an average travelling speed of interviewer.
- Time for interviewing $t_2(s) = m \cdot \bar{t}_H + n \cdot \bar{t}_P$ where m is number of dwellings taking part in survey, n is the number of individuals taking part in survey, $\bar{t}_H$ is an average time for a household interview, $\bar{t}_P$ is an average time for a personal interview.

The following cost function is used to measure the cost of the survey:

$$c(s) = t_1(s) + t_2(s) = \frac{\sum_{g=1}^{G} d_g}{\bar{v}} + m \cdot \bar{t}_H + n \cdot \bar{t}_P \tag{9}$$

6 Results

The results of the simulation study will be presented during the workshop.

References

European Communities (2003). *The european union labour force survey.* Office for Official Publications of the European Communities, Luxembourg.

Liberts, M. (2010). *Official statistics – methodology and applications in honour of Daniel Thorburn*, chap. The Redesign of Latvian Labour Force Survey. The Department of Statistics, Stockholm University (in collaboration with Statistics Sweden), Stockholm, Sweden, pp. 193–203.

Estimation in a mixed-mode, web and face-to-face, survey

Kaur Lumiste[1]

[1]University of Tartu, e-mail: kaur.lumiste@ut.ee

Abstract

Growing survey costs and falling response rates are problems for many survey companies and national statistics agencies. One heavily researched possible cure for this is to combine survey modes - mixed-mode design.

In September 2012 European Social Survey in Estonia, Slovenia and UK are planning an experiment with mixed-mode designs. Web and telephone survey modes are considered in conjunction with the usual face-to-face interview mode. The aim is to test for mode effects, influences in respondents' answers caused by the mixed-mode design, and develop means and protocols to avoid them.

In Estonia the experiment involves a web survey mode in conjunction with face-to-face interviews. At first an invitation is sent to sampled persons inviting them to fill the survey online. If a person shows no signs of activity, even after two reminders, then an interviewer is given a task to survey that person. The design gives us three random subgroups of the sample: people who fill the survey online, people who answer in the face-to-face interview and non-response subgroup. These subgroups tend to be different from one-another and, with auxiliary information, they can be used for better estimation.

Current paper gives a short overview on the preliminary studies made on estimation in this experimental mixed-mode design.

Keywords: Mixed-mode survey, web survey, face-to-face interview, estimation

1 Introduction

All survey designs pursue the somewhat incompatible objectives of reducing error and limiting costs. Survey companies and national statistics agencies are trying to find ways of making surveys more cost effective while not giving away precision. One heavily researched option is to combine survey modes (e.g. internet and postal survey or telephone and CAPI). Mixed-mode designs have special appeal for reducing coverage and non-response error, while also bringing costs down (Dillman & Messer, 2010). Ensuring that all members of a population have a known, nonzero chance of being sampled is very difficult, if not possible with certain designs. For example a web survey leaves out respondents who do not have access to internet, so a mixed-mode design should be considered.

Mixed-mode designs may reduce non-response error. People who are unwilling to participate in a telephone survey may be willing to respond by mail or over the internet. Groves & Kahn (1979) and Millar, Dillman, & O'Neill (2009) showed that some people prefer certain modes for being surveyed, while objecting to others. More importantly, those preferring different modes could differ from one another, as Link & Mokdad (2006) found that respondents to telephone and mail versions of the survey differed on demographic characteristics including gender, age, and income.

But all this does not come without drawbacks, using multiple modes may introduce mode effects, thereby increasing measurement error. For example Hochstim (1967) showed that personal interviews produced more "excellent" answers (40%) to a simple question, "Do you consider your health to be excellent, good, fair, or poor" than did mail surveys (30%). Interviewer presence encourages respondents to give answers consistent with social norms, a behaviour known as social desirability bias. Also different designs require different question wording, for example paper and web questionnaires may use check-all-that-apply questions, while telephone surveys use forced-choice items offering respondents a "yes/no" choice for each item. For a more complete list of mode effects the reader is referred to Dillman & Messer (2010).

In September 2012 three participating countries of the European Social Survey (ESS) will conduct a mixed-mode experiment in the background of data collection for ESS round 6. The experiment aims to test the feasibility of using other survey modes in conjunction with the face-to-face interviews used so far. Currently telephone and internet surveys are being considered. The experiment will be conducted simultaneously in Estonia, Great Britain and Slovenia, and Estonia will test CAPI in conjunction with the web survey method. First, sampled persons are invited to fill the ESS questionnaire online. If the invitation is ignored, as well as the two reminders, an interviewer is sent for a face-to-face interview.

The experiment's design divides sampled persons into two subgroups (with random sizes) - people who answer online and those who did not. The grouping is not completely random (like simple random sampling without replacement) since respondents' mode preference may be dependant on some demographic characteristics, as mentioned earlier. The subgroup of people who did not answer online is again divided into two groups - respondents by face-to-face interview, and non-respondents.

This paper presents preliminary studies on a possible estimation method in case of this special experimental case. First, two different estimators for the population total can be defined using these two groups of respondents and auxiliary information, and then also a linear combination of the two estimators is proposed.

2 Estimation in mixed-mode surveys

2.1 Preliminaries

Let $U = (1, 2, \dots, N)$ denote a finite population of N units. Let a random vector (design vector) $\mathbf{I} = (I_1, I_2, \dots, I_N)$ describe the sampling process on U and I_i is the sample inclusion indicator for unit $i \in U$. The probability sampling design generates for element i a known inclusion probability, $E(I_i) = \pi_i > 0$, and a corresponding sampling design weight $a_i = 1/\pi_i$. In case of non-response data can only be collected from a sample subgroup $r \subseteq s$. The study variable y is recorded for all $i \in r$ and our objective is to estimate the population total $Y = \sum_U y_i$. The basic design unbiased estimator of Y from a full sample s is $\hat{t}_{\text{HT}} = \sum_s a_i y_i$, the Horwitz-Thopmson (HT) estimator.

Auxiliary information has become more and more crucial in effective estimation and dealing with non-response. The auxiliary vector value $\mathbf{x}_i : J \times 1$ is assumed available for every element $i \in s$ (or every $i \in U$ if it is compiled from comprehensive registers) and J is the number of auxiliary variables available.

2.2 Mixed-mode

Since data from the respondents will be collected in two parts, let us define two random vectors:

$$\mathbf{W} = (W_1, W_2, \dots, W_n \mid s),$$

where $W_i = 1$ if unit i in sample s answers in the web mode and $W_i = 0$ otherwise, and

$$\mathbf{F} = (F_1, F_2, \dots, F_n \mid s),$$

where $F_i = 1$ if unit i in sample s answers in the face-to-face mode and $F_i = 0$ otherwise. Note that vector $\mathbf{W} + \mathbf{F}$ indicates the units that belong to set r. As the data collection begins the sample s is divided into two subsets - sampled persons who choose to answer online, $r_{web} = \{i \mid W_i = 1, s\}$, and remaining sampled persons i.e. $s_{ftf} = s - r_{web} = \{i \mid W_i = 0, s\}$.

We can now define $P(W_i = 1 \mid s) = p_i$, which is the probability that unit i will answer the survey via the internet, and an unbiased estimate for the population total Y can be found:

$$\hat{t}_{web} = \sum_{r_{web}} \frac{y_i}{\pi_i p_i},$$

since $\pi_i p_i = P(I_i = 1) \cdot P(W_i = 1 \mid s) = P(i \in s) \cdot P(i \in r_{web} \mid s) = P(i \in r_{web})$.

The probabilities p_i have to be estimated and this can be done using auxiliary information, but we will come back to this later in section 2.3.

Since sampled units, who do not answer online, are approached for a face-to-face interview, the probability of that happening is $1 - p_i$. We now define $P(F_i = 1 \mid s_{ftf}) = q_i$, which is the probability of person i answering to the survey in a face-to-face interview under the condition that he belongs to s_{ftf}. The following unbiased estimator can be constructed:

$$\hat{t}_{ftf} = \sum_{r_{ftf}} \frac{y_i}{\pi_i (1 - p_i) q_i}.$$

We now have two different estimates for Y and we can get a more efficient estimator by linearly combining them

$$\hat{t} = \alpha \hat{t}_{web} + (1 - \alpha) \hat{t}_{ftf}$$

where $\alpha \in (0,1)$ and can be found by minimizing $Var(\hat{t})$. In practice it would be very rare, but for simplicity let us assume that $\hat{t}_{web}$ and $\hat{t}_{ftf}$ are independent. Then the optimal α is

$$\alpha^* = \frac{Var(\hat{t}_{ftf})}{Var(\hat{t}_{web}) + Var(\hat{t}_{ftf})}.$$

2.3 Estimating mode participation probabilities

As mentioned earlier, research has shown that mode preference can be dependant on demographic characteristics like age and gender. Usually these variables can be retrieved from population registries for all sampled elements and can be taken as auxiliary variables. Särndal (2011) uses auxiliary information to estimate response probabilities $\theta_i = P(i \in r \mid s)$, but we adapt it to estimate mode participation probabilities p_i and q_i. In general case for estimating θ_i we need two conditions:

1. The estimates $\hat{\theta}_i$ for θ_i are linearly dependant on auxiliary variables $\mathbf{x}_i$, meaning that there is a constant vector $\boldsymbol{\lambda}$ so that

$$\hat{\theta}_i = \boldsymbol{\lambda}'\mathbf{x}_i. \tag{0.1}$$

2. The estimates $\hat{\theta}_i$ satisfy restrictions in the response/non-response case for I_i being here the response indicator:

$$\sum_s a_i(I_i - \hat{\theta}_i)\mathbf{x}_i = 0 \text{ or}$$

$$\sum_r a_i\mathbf{x}_i = \sum_s a_i\hat{\theta}_i\mathbf{x}_i. \tag{0.2}$$

With these restrictions $\boldsymbol{\lambda}$ can be found by substituting (0.1) into (0.2) so that we get

$$\sum_r a_i\mathbf{x}_i' = \boldsymbol{\lambda}'\sum_s a_i\mathbf{x}_i\mathbf{x}_i'.$$

By extracting $\boldsymbol{\lambda}'$and using it in (0.1), we get an estimate for the answering probability

$$\hat{\theta}_i = \left(\sum_r a_i\,\mathbf{x}_i\right)'\left(\sum_s a_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\mathbf{x}_i.$$

For mode participation probabilities p_i and q_i, given the auxiliary vector $\mathbf{x}_i$, the estimators take the following form:

$$\hat{p}_i = \left(\sum_{r_{web}} a_i\,\mathbf{x}_i\right)'\left(\sum_s a_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\mathbf{x}_i \text{ and } \hat{q}_i = \left(\sum_{r_{ftf}} a_i\,\mathbf{x}_i\right)'\left(\sum_{s_{ftf}} a_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\mathbf{x}_i.$$

3 Conclusions

Estonia's ESS team will conduct an experiment with combining web survey mode with CAPI. Estimators for such a design are studied and the preliminary results presented.

Further research aims to study the properties of these estimators, find the optimal α if the two population totals are correlated and test the estimators in a simulation study.

References

Dillman, D. A. & Messer, B. L. (2010). Mixed-Mode Survey. In: P. Marsden and J. Wrigth, ed. 2010 *Handbook of Survey Methodology*. Bingley: Emerald Publishing Limited, 551-574.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews.* NewYork: Acadmic Press.

Hochstim, J. R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976-989.

Link, M.W., & Mokdad, A. (2006). Can web and mail survey modes improve participation in an RDD-bsaed national health surveillance? *Journal of Official Statistics*, 22, 293-312.

Millar, M., Dillman, D. A. & O'Neill, A. C. (2009). *Are mode preferences real?* Technical Report 09-003, Social and Economic Sciences Research Center, Washington State University, Pullman, WA.

Särndal, C.-E., (2011). The 2010 Morris Hansen Lecture. Dealing with Survey Nonresponse in Data Collection, in Estimation. *Journal of Official Statistics* 28. 1-21.

Survey Sampling at Kyiv National Economic University

Tetiana Manzhos[1]

[1]Vadym Hetman Kyiv National Economic University, e-mail: tmanzhos@gmail.com

Abstract

The paper presents short review about Kyiv National Economic University and its history. It is given information about statistical disciplines for students of the University. Short description of a new course “Survey Sampling and Hypothesis Testing in Economics” is given.

Keywords: survey sampling theory, statistical disciplines.

1 KNEU: general information and brief historical overview

In 2011, our University celebrated its 105th anniversary. This year we will celebrate the 145th anniversary of our founder, a prominent historian Mytrofan Dovnar-Zapolsky (1867 - 1934). It was on his initiative that the Kyiv Graduate Commercial Courses were founded in Kyiv in 1906 as a private higher education establishment tasked with the training of human resources for the sectors of economy in the south of the Russian Empire. It became the second higher education establishment in the field of economics in the Empire and the first one within the territory of Ukraine. At that time, the Courses had 229 students and 22 teachers.

In 1908, Kyiv Graduate Commercial Courses were reorganized into Kyiv Commercial Institute and provided with own building. Commercial Institute was granted equal rights on a par with state-owned higher education establishments. As a result of strengthening of the role of economy in the society, the demand for the education in economics was growing. Accordingly, the number of students in the Institute was going up as well: the Institute had about 1,000 students and 34 professors and teachers by 1908. By 1914, there were 4,200 students and more than 50 teachers. In 1914, the construction of the 4th floor of the Institute's building was completed. However, World War I started the same year and the hostilities lasted till 1921 in view of the struggle over the Ukrainian-populated lands of neighboring states, which produced mainly negative impact on the education sector. Kyiv Commercial Institute was evacuated in autumn 1915 to Saratov. In summer 1916 it was returned to Kyiv.

During the Freedom Movement in 1917 to 1921, the Institute was also involved into the revival of the Ukrainian state: one of its graduates (M.M. Kovalevsky) was a minister in the government of the Ukrainian People's Republic. After the final imposition of the Soviet regime in Ukraine, the Institute was conveyed into the ownership of the state and changed its name. Between 1920 and 1930 it was named Kyiv National Economy Institute.

On 1 October 1930, Kyiv National Economy Institute was transformed into two institutes: Kyiv Exchange and Distribution Institute and Kyiv Finance and Economics Institute. In 1931, Kyiv Exchange and Distribution Institute was liquidated. Kyiv Finance and Economics Institute was moved to Kharkiv in 1934 and stayed there till 1941 having changed its name to Kharkiv (Ukrainian) Finance and Economics Institute. During the Kharkiv period, there were about 700 students and 45 teachers in the Institute. Up to 400 students and teachers joined the Red Army, when the Great Patriotic War started. The rest of employees and students were evacuated to local finance and economics institutes in Irkutsk and Tashkent. At the end of the war (1944) the Institute was permitted to return to Kyiv and resumed working in its native city as Kyiv Finance and Economics Institute.

Structurally, the Institute consisted of two faculties after 1945: the Finance Faculty with the finance and credit departments and the Planning Faculty with departments of industrial economics and planning, and agricultural economics and planning. The duration of studies was 4 years. In 1946, the post-graduate school of the Institute was restored. The number of students kept growing. The main building of the Institute was completed in 1958. It permitted normalizing the instructional process and increased the number of students (from 343 in 1945 to 5,000 in 1960).

Since the Institute has started training specialists in late 1950s in 12 specialities going beyond the finance and economics profile, the Ministry of Higher and Specialized Secondary Education of the Ukrainian RSR made a decision in 1960 to rename Kyiv Finance and Economics Institute into Kyiv National Economy Institute, thus broadening the range of specialities, in which the Institute trained specialists. At that time, the Institute had 5 Faculties: Industrial Economics, Agricultural Economics, Finance and Economics, Accounting and Economics, and Economics and Statistics.

Upon restoration of Ukraine's independence, Kyiv National Economy Institute made a lot to upgrade the system and the contents of the economic education, to improve the training of specialists for various sectors of the national economy of our country, and to develop appropriate scientific and educational literature. In addition to well-established international relations with ex-Communist countries, close relations were set up with economics higher education institutions of Austria, the UK, Germany, the Netherlands, the USA, France and other leading countries of the world.

Achievements of the Institute in the development of Ukraine's economy were recognized at the state level. Resolution of the Cabinet of Ministers of Ukraine of 25 August 1992 transformed Kyiv National Economy Institute into Kyiv State Economics University. On 27 February 1997, the President of Ukraine granted Kyiv State Economics University the status of a national university in recognition of its thorough work focused on training highly qualified specialists for various branches of economy of our state and the development of the domestic economic science.

In 2005, Kyiv National Economics University was named after V.P. Hetman, a prominent Ukrainian economist and the founder of the domestic currency of Ukraine, the builder of its banking sector. He obtained education in Economics in our Institute. The 100th anniversary of the University was celebrated in November 2006 at the national level.

Now Vadym Hetman Kyiv National Economics University consists of nine Faculties: Economics and Management, International Economy and Management, Law, Human Resources Management and Marketing, Accounting and Economics, Agro industrial Sector Economics, Finance and Economics, Credit and Economics, and Information Systems and Technologies. The University also includes the Post-graduate Education Centre, the Master Training Centre, the Pre-university Training Department, the Instruction Centre, the Instruction Methodology Unit, the post-graduate and doctoral schools, Kryvyi Rih and Crimean Institutes of Economics, target-oriented lyceums and colleges, library and computer centers, a museum, a publishing office, etc. There are more than 38,000 students.

2 Mathematical statistics as a discipline for future economists

Mathematics, as educational discipline, took the important place in all of the economic specialities. Wide application of mathematical methods is a feature of modern economy. The Department of Higher Mathematics conducts the fundamental mathematical training of students of all faculties, except the Faculty of Law. In the first year of studies, students study such courses, as “Mathematics for economists” and “Theory of probability and mathematical statistics”. The purpose of course “Theory of probability and mathematical statistics” is to acquaint students with basic concepts, methods, theorems and formulas of probability theory and mathematical statistics and help them to get primary skills in solving different problems. This course is for students of the first year of training and its studying continues during second semester.

Course “Theory of probability and mathematical statistics” consists of main concepts of probability theory and such topics of mathematical statistics as descriptive statistics (graphs and measures), confidence intervals (partially), hypothesis testing (partially), simple linear regression. Fundamental theoretical knowledge and skills of using probabilistic and statistical apparatus for different economic researches, analysis and predicting are the basis for successful learning of main economic disciplines such as macroeconomics, microeconomics, financial analysis, mathematical modeling etc. That’s why Department of Higher Mathematics is preparing new special statistical courses for our students. Another reason for such creation is that the main course of mathematical statistics contains not all topics which can be needed by our students for their future scientific researches and working. Some topics such as testing of statistical hypotheses are being studied not enough and survey sampling is not involved in main course of statistics at all. But now our university is successfully integrating into the European system of education based on the Bologna declaration, which defined the approaches to creating a single European educational environment. Therefore teachers of Department of Higher Mathematics obtained an opportunity to create new useful for students and up to date courses.

Main problem of improving studying process of mathematical statistics disciplines in our university is shortage of statistical literature in Ukrainian for students of economic spesialities without strong mathematical background.

3 Course of Survey Sampling at the Economic University

Recently special course named “Survey Sampling and Hypothesis Testing in Economics” was created by teachers of our department. Despite some bureaucratic problems we obtained official permission to teach this course. Students of the first year of training have to choose some part of special courses for learning during the second year. Now created course is one of them. Next academic year this course will be carried out for the first time.

This course has been developed for all specialities except Law specality and it contains of 5 ECTS Credits. It will be divided in two parts – “Survey Sampling” (with 28 hours of lectures and 24 hours of practical lessons) and “Testing of Statistical Hypothesis” (with 26 hours of lectures and 24 hours of practical lessons). This course is oriented on future economists who have no strong mathematical background. Students of economic spesialities need more examples and less complicated theoretical material. They are interested how it works in practice. It was taken into account when new course was developing.

A short program of the firs part of the course is:

- Goals and applications of survey sampling in economic researches, main concepts and definition
- Simple random sampling with and without replacement
- Estimators of total, mean, proportion in the population

- Confidence intervals and sample size
- Systematic sampling
- Stratified sampling
- Ratio and regression estimators
- Single-stage and multistage cluster sampling
- Errors in surveys, their sources and methods of reduction

Second part of the course consists of testing of parametric and nonparametric hypotheses.

On practical lessons of "Sample surveys and hypothesis testing in economics" students will solve different exercises with using software such as MS Excel, Statistica, Wolfram Mathematica. Two individual tasks based on exercises of S. L. Lohr were involved in the plan of practical work too. For creation a program and lectures of the course mainly books listed in the references were used.

References

Lohr, S. (1999). *Sampling: Design and Analysis.* Duxbury Press, Pacific Grove.

Kvanli, A., Pavur, R., & Keeling, K. (2003) *Introduction to Business Statistics: A Computer Integrated, Data Analysis Approach.* South-Western Publishing.

Vasylyk, O. & Yakovenko, T. (2010). *Lectures on Survey Sampling Theory and Methods*. Kyiv National University (in Ukrainian)

Chernyak, O. (2001). *Survey Sampling Technique.* Kyiv (in Ukrainian)

Estimation of social and economic characteristics of immigrants

Inga Masiulaitytė-Šukevič[1]

[1]Statistics Lithuania, Mykolas Romeris University,
e-mail: inga.masiulaityte@stat.gov.lt

Abstract

International migration is widely spread in Europe. A lot of persons are migrating abroad for finding better life: new job, new studies, or new possibilities. Economic crises in Lithuania consequences a large emigration and small immigration flows. Migration is a form of geographic mobility involving a change of usual residence between clearly defined geographic units. The main source of data on international migration in Lithuania is the central database of the Residents' Register of the Residents' Register Service under the Ministry of the Interior of the Republic of Lithuania. This study surveys how the immigrants are living in Lithuania. Using population of immigrants and the EU Statistics on Income and Living Conditions (EU-SILC) survey, estimation of social and economic characteristics of immigrants will be done.

Keywords: international migration, immigrant

References

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

Siegel, J.S.,Swanson, D. A.Wickham, H. (2008). *Methods. Materials. Demography*. Emerald Group Publishing Limited.

Lack of Balance Indicator for Data Collection

Maiken Mätik[1]

[1]University of Tartu, e-mail: maiken.matik@gmail.com

Abstract

The purpose of this paper is to study novel tools that measure balance of the response set against the full sample with respect to auxiliary variables. A measure called "lack of balance" is introduced. Its statistical properties are explored and an instrument called "balance indicator" is defined. Illustrative examples about the special cases of the balance indicator and related matters are given. A practical experiment was carried out on real data to illustrate theory about balance indicators. The experiment confirmed that balance indicator really shows balance under random or independent nonresponse and imbalance under dependent (on auxiliary variables) nonresponse.

Keywords: Auxiliary information, balance indicator, balanced response set, lack of balance

1 Introduction

The purpose of survey sampling is to give information about unknown parameters in the population $U = \{1, \ldots, N\}$. Depending on the purpose and scope of the survey, special sampling design is used in U. With the design, inclusion probabilities, weights and other design characteristics are defined. For every object $k \in U$ we have positive inclusion probabilities $\pi_k = P(k \in s) > 0$ and for every object $k \in s$ we have a design weight $d_k = 1/\pi_k$.

Nowadays, nonresponse is a very common issue in survey sampling. There are always objects, from whom information is not received. We refer to the response set with symbol r. For example, many people with higher salary will not give their income data which leads to imbalanced response set with respect to the full sample. Survey estimates from respondents will then have nonresponse bias. Special efforts should be made already at the data collection stage to measure nonresponse effect, and possibly to reduce this effect. In this paper we introduce and study the tools given in Särndal (2011), and explored in Mätik (2012).

2 Response Rate and Response Probabilities

Lets assume we have the probability sample s with size n which means we have the objects with some auxiliary information that we gathered from the registers. But only a subset r with size m from s responds. Response rate is defined as

$$P = \frac{\sum_r d_k}{\sum_s d_k}. \tag{1}$$

We see that for equal d_k, $P = m/n$. The response indicator I is the binary random variable, observed for $k \in s$, with value $I_k = 1$ for $k \in r$ and $I_k = 0$ for $k \in (s - r)$.

Definition 2.1 The response probability for object $k \in s$ is defined through response indicator in a following way,

$$E(I_k|s) = P(I_k = 1|s) = \theta_k. \tag{2}$$

Response probabilities for all $k \in s$ are unknown parameters.

3 Measuring Lack of Balance

Assume we know a $J-$dimensional auxiliary variable vector $\mathbf{x}_k$ for each $k \in s$.

Definition 3.1 We call the response set r balanced when the means for appropriate auxiliary variables in r equal to corresponding means in the sample s.

We consider auxiliary vectors $\mathbf{x}_k$ which for some constant vector $\boldsymbol{\mu} \neq 0$, satisfy

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \qquad \text{for all} \quad k \in U. \tag{3}$$

We define two J-dimensional mean vectors and two computable $J \times J$ non-singular weighting matrices:

$$\bar{\mathbf{x}}_{r;d} = \sum\nolimits_r d_k \mathbf{x}_k \Big/ \sum\nolimits_r d_k, \tag{4}$$

$$\boldsymbol{\Sigma}_r = \sum\nolimits_r d_k \mathbf{x}_k \mathbf{x}_k^{'} \Big/ \sum\nolimits_r d_k, \tag{5}$$

$$\bar{\mathbf{x}}_{s;d} = \sum\nolimits_s d_k \mathbf{x}_k \Big/ \sum\nolimits_s d_k, \tag{6}$$

$$\boldsymbol{\Sigma}_s = \sum\nolimits_s d_k \mathbf{x}_k \mathbf{x}_k^{'} \Big/ \sum\nolimits_s d_k. \tag{7}$$

Auxiliary vectors that satisfy (3) also satisfy on all outcomes (s,r):

$$\bar{\mathbf{x}}_{r;d}^{'} \boldsymbol{\Sigma}_r^{-1} \bar{\mathbf{x}}_{r;d} = \bar{\mathbf{x}}_{r;d}^{'} \boldsymbol{\Sigma}_r^{-1} \bar{\mathbf{x}}_{s;d} = \bar{\mathbf{x}}_{r;d}' \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_{s;d} = \bar{\mathbf{x}}_{s;d}^{'} \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_{s;d} = 1. \tag{8}$$

Definition 3.2 A measure

$$\boldsymbol{D}^{'} \boldsymbol{\Sigma}_s^{-1} \boldsymbol{D} = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})^{'} \boldsymbol{\Sigma}_s^{-1} (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d}), \tag{9}$$

is defined as lack of balance indicator. It is a quadratic form in the differences in auxiliary variable means between the response set and the whole sample.

The lack of balance indicator refers to balance when the auxiliary variable means between the response set and the whole sample are equal, then $\mathbf{D} = 0$ and $\mathbf{D}^{'} \boldsymbol{\Sigma}_s^{-1} \mathbf{D} = 0$.
For one dimensional auxiliary vector $\mathbf{x}_k = \mathrm{x}_k$, the lack of balance indicator is

$$\mathbf{D}^{'} \boldsymbol{\Sigma}_s^{-1} \mathbf{D} = (\bar{\mathrm{x}}_{r;d} - \bar{\mathrm{x}}_{s;d})^2 \cdot \frac{\sum_s d_k}{\sum_s d_k \mathrm{x}_k^2}.$$

4 Estimated Response Probabilities

Looking for an estimator of θ_k, linearly depending on $\mathbf{x}_k$,

$$\hat{\theta}_k = \boldsymbol{\lambda}^{'} \mathbf{x}_k, \tag{10}$$

one gets,

$$\hat{\theta}_k = t_k = (\sum\nolimits_r d_k \mathbf{x}_k)^{'} (\sum\nolimits_s d_k \mathbf{x}_k \mathbf{x}_k^{'})^{-1} \mathbf{x}_k. \tag{11}$$

The mean over r, and the mean and variance over s of the estimated response probabilities t_k are now related to the response rate P and lack of balance indicator in the following way:

$$\bar{t}_{r;d} = P \times \bar{\mathbf{x}}_{r;d}^{'} \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_{r;d}, \tag{12}$$

$$\bar{t}_{s;d} = P, \tag{13}$$

$$S^2_{t|s;d} = \bar{t}_{s;d}(\bar{t}_{r;d} - \bar{t}_{s;d}) = P^2 \times \mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D}. \tag{14}$$

For constant response probability estimates, $\hat{\theta}_k = t_k = c$, the variance of the estimates is zero. Consequently, for $P \neq 0$ the lack of balance indicator is zero for $t_k = c$. Thus, for constant response probabilities the response set r is always balanced and represents the whole sample s.

We see that (13) and (14) now define the lack of balance indicator as the coefficient of variation of estimated response probabilities,

$$cv_{t|s;d} = \frac{S_{t|s;d}}{\bar{t}_{s;d}} = \frac{\sqrt{P^2 \times \mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D}}}{P} = (\mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D})^{1/2}.$$

The upper bound of the lack of balance indicator is

$$\mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D} \leq Q - 1,$$

where Q is inverse value of response rate P. We call $Q - 1$ nonresponse odds.

5 Balance Indicators

We consider three types of the balance indicators, all of them measured on the unit interval scale:

$$BI_1 = 1 - \frac{\mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D}}{Q-1} = 1 - \frac{S^2_{t|s;d}}{P(1-P)}, \tag{15}$$

$$BI_2 = 1 - 4P^2\mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D} = 1 - 4S^2_{t|s;d}, \tag{16}$$

$$BI_3 = 1 - 2P(\mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D})^{1/2} = 1 - 2S_{t|s;d}. \tag{17}$$

For every outcome (s, r) and a fixed auxiliary vector $\mathbf{x}_k$ we have

$$0 \leq BI_1 \leq BI_2 \leq 1 \quad \text{ja} \quad 0 \leq BI_3 \leq BI_2 \leq 1.$$

These indexes show complete imbalance with the value 0, and complete balance with the value 1. It is important to remember, that balance/imbalance is measured with respect to chosen auxiliary vector.

6 Simulation Example

In this simulation example we used data about Estonian health care employees. There were 21761 objects in the register and 29 variables were measured for each individual. In our experiment we used one categorical variable, *education* (5 categories), and one continuous variable, *age*.

In the first part of the experiment we considered a response set that was independent form any of the variables. Both, the sample s (with size $n = 1000$) and the response set r (with size $m = 700$) were drawn with simple random sampling. Thus, the response rate was $P = 0.7$. The theoretical response probabilities were equal for all $k \in s$, so $\theta_k = m/n = 0.7$. We calculated the estimated response probabilities using three auxiliary vectors $\mathbf{x}_k$, extended stepwise. The results are shown in Table 1. The estimates t_k had very small variation around their mean 0.7 which equals theoretical θ_k. The calculated balance indicators approve theory that for independent from the variables response, the response set is balanced and represents the whole sample.

Table 1: Independent nonresponse

Auxiliary vector $\mathbf{x}_k$	Estimates t_k in sample s			
	mean	sd	BI_1	BI_2
One education category	0.7	0.0020	1.0000	1.0000
Four education categories	0.7	0.0103	0.9995	0.9996
Four education categories and age	0.7	0.0292	0.9959	0.9966

In the second part of the simulation exercise we drew a simple random sample s (with size $n = 1000$) but the response set r (with size $m = 700$) was generated as dependent on the variable *age*. Older people had bigger response probability. Thus our response set is imbalanced and the balance indicators should approve it. Again, we calculated the estimated response probabilities using three auxiliary vectors $\mathbf{x}_k$ built step by step. The results are shown in Table 2. The mean of t_k is still 0.7 but their variability is now bigger. For the first two $\mathbf{x}_k$ vectors, the indicators show balance because the response was not dependent on the variable *education*. For the third auxiliary vector, that includes *age*, the indicators approve that the response set is imbalanced.

Table 2: Dependent nonresponse

Auxiliary vector $\mathbf{x}_k$	Estimates t_k in sample s			
	mean	sd	BI_1	BI_2
One education category	0.7	0.0260	0.9968	0.9973
Four education categories	0.7	0.0272	0.9965	0.9970
Four education categories and age	0.7	0.1871	0.8333	0.8600

The experiment confirmed that balance indicators show balance under random or independent nonresponse. They show imbalance if the variables related to the response mechanism are included in $\mathbf{x}_k$.

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): 1-21.

Mätik, M., 2012. Dealing with Survey Nonresponse in Data Collection and in Estimation. Bachelor thesis (in Estonian). University of Tartu.

Estimation strategy for small areas, a case study

Nekrašaitė-Liegė Vilma[1]

[1]Vilnius Gediminas technical university, e-mail: nekrasaite.vilma@gmail.com

Abstract

The purpose of this research is to find optimal strategy (pair of sample design and estimator) for small area estimation. Thus, the definition of a balanced sample and two special cases of balanced samples are presented. The study variable and auxiliary information are time series, thus the different unit level panel-type models are used not only in estimation stage, but and in sample selection stage. The simulation results showed, that the impact of the chosen model is larger for the small domains than for the large ones. Also results showed that the use of the panel type model in sample selection and estimation stages improves the accuracy of the estimate.

Keywords: small area, balanced sample, panel-type model.

1 Introduction

As mentioned by Ghosh & Rao (1994) the term "small area" and "local area" are commonly used to denote a small geographical area, such as a county, a municipality or a census division. They may also describe a "small domain", i.e., a small subpopulation such as a specific age-sex-race group of people within a large geographical area.

The focus on small area estimation (SAE) is made because the demand for data at lower geographic levels is always present, especially from local governments and from businesses needing to make investment, marketing, and location decisions that depend on knowledge of local areas.

The small area problem is usually considered to be treated via estimation (Ghosh & Rao, 1994). However, if the domain indicator variables are available for each unit in the population there are opportunities to be exploited at the survey design stage. Thus in this paper I am interesting in an overall strategy that deals with small area problems, involving both planning sample design and estimation aspects.

2 Main notations

A finite *population* $U = \{u_1, u_2, ..., u_N\}$ of the size N is considered. For simplicity, in the sequel we identify a population element u_k and its index k. Hence $U = \{1, 2, ..., N\}$.

The elements k $(k = 1, \ldots, N)$ of the population U has two components $y(t)$ and $\mathbf{x}(t)$. The values of these components depends on time. The component $y(t)$ defines the value of a *study variable* (variable of interest) in time t, and the component $\mathbf{x}(t) = \{x_1(t), x_2(t), \ldots, x_J(t)\} \in \mathbb{R}^J$ defines the values of the J *auxiliary variables* in time t.

The population is divided into D nonoverlapping *domains* (subpopulations) $U^{(d)}$ of size $N^{(d)}$, where $d = 1, \ldots, D$. Domain indicator variables define whether $k \in U$ belongs to a given domain:

$$q_k^{(d)} = \begin{cases} 1, & \text{if } k \in U^{(d)}, \\ 0, & \text{otherwise,} \end{cases} \qquad \forall k \in U, d = 1, \ldots, D. \tag{1}$$

The *parameter of interest* is a *domain total* in time moment t:

$$TOT^{(d)}(t) = \sum_{k \in U^{(d)}} y_k(t) = \sum_{k \in U} q_k^{(d)} y_k(t), \quad d = 1, \ldots, D; \quad t = 1, 2, \ldots \tag{2}$$

To estimate $TOT^{(d)}(t)$, we need information about unknown variable $y(t)$. This information is collected by sampling. The *sampling vector*

$$\underline{\mathbf{S}}(t) = (\underline{S}_1(t), \underline{S}_2(t), \dots, \underline{S}_N(t)) \tag{3}$$

is a random vector whose elements $\underline{S}_k(t)$ indicate the number selections for k in time t. The distribution of $\underline{\mathbf{S}}(t)$, denoted by $p(.)$, is called a *sample design.* The realization $\mathbf{S}(t) = (S_1(t), S_2(t), \dots, S_N(t))$ of $\underline{S}_k(t)$ is called *sample.* It define the *sample set* $s(t) = \{k : k \in U, S_k(t) \geq 1\}$.

3 Balanced samples

The sample might be balanced or unbalanced. A sample is said to be balanced if, for a vector of auxiliary variable $\mathbf{z}(t) = \{z_1(t), z_2(t), \dots, z_L(t)\} \in \mathbb{R}^L$,

$$\sum_{k \in s(t)} \frac{\mathbf{z}_k(t)}{\pi_k(t)} = \sum_{k \in U} \mathbf{z}_k(t). \tag{4}$$

In other words, in a balanced sample, the total of the z-variables are estimated without error. Let us note that two different sets of variables have been introduced in order to underline that the set of variables available at the design stage ($\mathbf{z}$ variables) could be different from the set available at the estimation stage ($\mathbf{x}$ variables) even if in many practical situations they could be the same. Here, an element's k inclusion probability in time t is denoted as $\pi_k(t)$.

Almost all the other sampling techniques are particular cases of balanced sampling. Some well-known sampling designs are particular cases of balanced sampling:

1. Sampling with a fixed sample size is a particular case of balanced sampling. In this case, the only balancing variable is $\pi_k(t)$. The balancing equations given in (4) become

$$\sum_{k \in s} \frac{\pi_k}{\pi_k} = \sum_{k \in s} 1 = \sum_{k \in U} \pi_k,$$

which means that the sample size must be fixed.

2. Stratification is a particular case of balanced sampling. In this case, the balancing variables are the indicator variables of the strata

$$\delta_{kh} = \begin{cases} 1, & \text{if } k \in U_h, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where U_h denotes population part, which belongs to h, $h = 1, ..., H$ strata. Since the inclusion probabilities in stratum h are $\pi_k = n_h/N_h$, $k \in U_h$, the balancing equations become

$$\sum_{k \in s} \frac{N_h \delta_{kh}}{n_h} = \sum_{k \in U} \delta_{kh} = N_h, \quad h = 1, ..., H,$$

and are exactly satisfied.

The main reason why a balanced sample is used in my research is that the Deville & Tillé (2004) showed, that the optimal strategy contains a balanced sample. Thus, to select a balanced sample is not easy. The one of the quickest way to select balanced sample is to use cube method.

3.1 Cube method

The algorithm of the cube method was proposed by Deville & Tillé (1998), and the method was published in Tillé (2001) and Deville & Tillé (2004). This method is general in the sense that the inclusion probabilities are exactly satisfied, that these probabilities may be equal or unequal and that the sample is as balanced as possible.

It is possible to get SAS/IML version of Cube method done by Chauvet & Tillé (2006) and it is also available on the University of Neuchatel Web site. This software program is free, available over the Internet and is easy to use.

4 Model-based sample design

In this paper balanced samples are selected not only using well known simple random sample, stratified simple random sample, but also using model-based sample (Nekrašaitė-Liegė *et al.*, 2011). The suggested model-based sample design consists of three steps:

1. Model construction and estimation of it's coefficients;
2. Estimation of the variance of the prediction error;
3. Construction of the sample design $p(.)$.

In the first step the model is fitted to the available auxiliary data. In the second step the prediction errors (residuals)

$$\hat{\varepsilon}_k(t) = \hat{y}_k(t) - y_k(t), \quad t \in \mathcal{T}_k \in \{1, 2, ..., T.\}, \tag{6}$$

are calculated and the variance of prediction error in each domain, $\sigma^{(d)2}$, is estimated. Finally, in the third step the (approximately) optimal sample design $p(.)$ (actually, the Neyman stratified simple random sample, (Särndal *et al.*, 2003)) based on the estimated variances $\widehat{\sigma^{(d)2}}$ is constructed. Thus, the less model-based prediction accuracy in the domain the more elements from this domain are drawn.

5 Estimators and models

After the sample is selected the domain total is calculated using GREG-type (Lehtonen *et al.*, 2003) estimator:

$$\widehat{TOT}^{(d)}_{GREG}(t) = \sum_{k \in U^{(d)}} \hat{y}_k(t) + \sum_{k \in s(t) \cap U^{(d)}} 1/\pi_k(t)(y_k(t) - \hat{y}_k(t)). \tag{7}$$

where $\hat{y}_k(t)$ denotes the prediction of $y_k(t)$ under the assumed super population model. The predictions $\{\hat{y}_k(k); k \in U\}$ differ from one model specification to another, depending on the functional form and from the choice of the auxiliary variables.
In this paper a general panel data model with random effects is considerate as working super population model:

$$\begin{aligned} Y_k(t) =& \beta_{0,g(k)}(t) + r_{0,k}(t) + \sum_{j=1}^{J} [\beta_{j,g(k)}(t) + r_{j,k}(t)] X_{j,k}(t) + \\ &+ \sum_{i=1}^{m} \alpha_{i,g(k)} \mu_i(t) + \varepsilon_k(t), \quad k \in U. \end{aligned} \tag{8}$$

Here $X_{j,k}(t)$, $j = 1, 2, ..., J$, are fixed-effects variables, $\beta_{0,g(k)}(t)$, $\beta_{1,g(k)}(t)$, ..., $\beta_{J,g(k)}(t)$ are the unknown fixed-effects model coefficients, which are the same in group $g(k)$. The groups $g(k)$ divides population U into G nonoverlaping groups which in some special cases can be the same as domains d, $d = 1, ..., D$. The unknown random-effects models coefficients are denoted as $r_{0,k}(t)$, $r_{1,k}(t)$,..., $r_{J,k}(t)$ ($r_{p,k}(t) \sim IID(0, \lambda^2_{0,g(k)}(t)), g(k) = 1(k), ..., G(k), p = 0, ..., J$. The model error is denoted as $\varepsilon_k(t)$ ($E_M(\varepsilon_k(t)) = 0$, $VAR_M(\varepsilon_k(t)) = \nu_k^2\sigma^2$, $\forall k \in U$ and $cov(\varepsilon_k(t), \varepsilon_l(t)) = 0$ when $k \neq l$). It should be noticed that model error $\varepsilon_k(t)$ and the random-effects model coefficients $r_{0,k}(t)$, $r_{1,k}(t)$,..., $r_{J,k}(t)$ are conditionally independent if values of $X_{j,k}(t)$, $j = 1, 2, ..., J$, are known. The component $\sum_{i=1}^{m} \alpha_{i,g(k)} \mu_i(t)$ represents a time trend. The structure of this component depends on historical auxiliary information and is specified using exploratory analysis.
Below several special cases of this model are described:

- *Example 1.* Let $\beta_{0,g(k)}(t) = \beta_0(t)$, $r_{0,k}(t) = 0$, $\beta_{j,g(k)}(t) = \beta_j(t)$, $r_{j,k}(t) = 0$, $j = 1, ..., J$ and t is equal to one moment (let this moment is notated as W). Then the generalized unit level model has such form
$$Y_k(W) = \beta_0(W) + \sum_{j=1}^{J} \beta_j(W)X_{j,k}(W) + \varepsilon_k(W), \quad k \in U. \tag{9}$$
This model is known as common model (Lehtonen *et al.* (2003)), because it has the same model parameters for all domains.

- *Example 2.* Let $\beta_{0,g(k)}(t) = \beta_0^{(d)}(t)$, $r_{0,k}(t) = 0$, $\beta_{j,g(k)}(t) = \beta_j(t)$, $r_{j,k}(t) = 0$, $j = 1, ..., J$ and t is equal to one moment (let this moment is notated as W). Then the generalized unit level model has such form
$$Y_k(W) = \beta_0^{(d)}(W) + \sum_{j=1}^{J} \beta_j(W)X_{j,k}(W) + \varepsilon_k(W), \quad k \in U. \tag{10}$$
This model is known as model with domain-intercept (Lehtonen *et al.* (2003)), because it has the same slopes but separate intercepts for all domains.

- *Example 3.* Let $\beta_{0,g(k)}(t) = \beta_{0,g(k)}$, $r_{0,k}(t) = 0$, $\beta_{j,g(k)}(t) = \beta_{j,g(k)}$ $r_{j,k}(t) = 0$, $j = 1, ..., J$. Then the generalized unit level model has such form
$$Y_k(t) = \beta_{0,g(k)} + \sum_{j=1}^{J} \beta_{j,g(k)}X_{j,k}(t) + \varepsilon_k(t), \quad k \in U. \tag{11}$$
This model is fixed effect panel data model. Here models coefficients $\beta_{0,g(k)}$, $\beta_{1,g(k)}$, ..., $\beta_{J,g(k)}$ do not depend on time which means they are the same for the all periods of time. Such model is very useful in practice since it enables one to find model coefficients just using data from the past. The current data might be use just for prediction.

- *Example 4.* Let $\beta_{0,g(k)}(t) = \beta_0(t)$, $r_{0,k}(t) = r_{0,g(k)}(t)$, $\beta_{j,g(k)}(t) = \beta_j(t)$, $r_{j,k}(t) = 0$, $j = 1, ..., J$ and t is equal to one moment (let this moment is notated as W). Then the generalized unit level model has such form
$$Y_k(W) = \beta_0(W) + r_{0,g(k)}(W) + \sum_{j=1}^{J} \beta_j(W)X_{j,k}(W) + \varepsilon_k(W), \quad k \in U. \tag{12}$$
This model is known as mixed model with random-intercept, because it has the same fixed parameters for all domains. The random effect is defined at the group level.

- *Example 5.* Let $\beta_{0,g(k)}(t) = \beta_0^{(d)}$, $r_{0,k}(t) = r_{0,g(k)}$, $\beta_{j,g(k)}(t) = \beta_j^{(d)}$, $r_{j,k}(t) = 0$, $j = 1, ..., J$. Then the generalized unit level model has such form
$$Y_k(t) = \beta_0^{(d)} + r_{0,g(k)} + \sum_{j=1}^{J} \beta_j^{(d)}X_{j,k}(t) + \varepsilon_k(t), \quad k \in U. \tag{13}$$
It is assumed that the model coefficients $\beta_0^{(d)}$, $\beta_j^{(d)}$, $j = 1, ..., J$, and the random effects $r_{0,g(k)}$ do not depend on time (they are the same during the different time periods).

- *Example 6.* Let $\beta_{0,g(k)}(t) = \beta_0^{(d)}$, $r_{0,k}(t) = r_{0,g(k)}$, $\beta_{j,g(k)}(t) = \beta_j^{(d)}$, $r_{j,k}(t) = 0$, $j = 1, ..., J$. Then the generalized unit level model has such form
$$Y_k(t) = \beta_0^{(d)} + r_{0,g(k)} + a_0^{(d)}t + \mathbf{a}'^{(d)}\alpha(t) + \sum_{j=1}^{J} \beta_j^{(d)}x_{j,k}(t) + \varepsilon_k(t), \quad k \in U. \tag{14}$$
This is a panel data model with a linear trend and a seasonal component, $\mathbf{a}^{(d)\prime}\alpha(t)$, $\mathbf{a}^{(d)} \in \mathbf{R}^3$.

6 Simulation and Conclusions

For the simulation experiment, a real population from Statistics Lithuania is used. Enterprisers which are responsible for education are taken as the finite population. Information about these enterprisers is taken 20 times (each quarter from 2005 till 2009). The average number of enterprises in each quarter is 750 (Number of population).
The study variable $y_k(t)$ is the income of an enterprise k and the auxiliary variables are the number of employers $x_{1,k}(t)$, tax of value added (VAT) $x_{2,k}t$ and various indicators (specification of enterprise (5 indicators), size of enterprise (3 indicators), region (6 indicators)) $x_{j,k}$, $j = 3, ..., 15$.
The total income in a domain in each quarter in 2008 and 2009 is chosen as the parameter of interest ($T + l$, $T = 12$, $l = 1, ..., 8$). The domain is chosen as counties (there are 10 counties in Lithuania) and specification of enterpriser (5 specifications). So, in this research the study variables are elements of a time series with 8 elements and the total number of domains of interest is 120. The number of enterprises in each domain varies from 6 to over than 300.
The comparison of results of 1000 simulations using different models and auxiliary information in both stages (sample selection and estimation) showed that the impact of the model is larger for the small domains than for the large domains. Also results showed that the use of the panel type model in sample selection and estimation stages improves the accuracy of the estimate. More results and conclusions will be presented during presentation.

References

Chauvet, G. & Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics* **21**, 9 – 31.

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

Deville, J.-C. & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika* **91**, 893 – 912.

Ghosh, M. & Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science* **9**, 55 – 93.

Lehtonen, R., Särndal, C.-E. & Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology* **29**, 33 – 44.

Nekrašaitė-Liegė, V., Radavičius, M. & Rudys, T. (2011). Model-based design in small area estimation. *Lithuanian mathematical journal* **51**, 417 – 424.

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

Tillé, Y. (2001). *Thorie des sondages : chantillonnage et estimation en populations finies.* Dunod, Paris.

Indicators with respect to process model for editing in Statistic Finland

Pauli Ollila[1], Outi Ahti-Miettinen[2] and Saara Oinonen[3]

[1]Statistics Finland, e-mail: pauli.ollila@stat.fi
[2]Statistics Finland, e-mail: outi.ahti-miettinen@stat.fi
[3]Statistics Finland, e-mail: saara.oinonen@stat.fi

Abstract

Indicators developed for editing models have two main functions. On one hand, they are used to control error identification and correction actions. On the other hand, indicators can analyse effect of the editing actions on the quality of data at the different stages of the editing model and estimate the overall quality of the final data. We divide indicators into three groups: Indicators of raw data, indicators that relate to the error identification and indicators that relate to error correction. In this paper, we discuss what kind of indicators we need and in which stages of the editing model they should be computed. We also make an overview of the demands of the ESS standard of quality reporting in editing and outline recommendations for what indicators to use.

Keywords: Editing; Imputation; Indicator; Process model

1 Introduction

Statistics Finland has carried out an editing project, whose main task was to survey editing and imputation practices at Statistics Finland and produce Editing Model for Statistics Finland. Statistical editing refers to activities by which statistical data are checked to detect of missing, invalid or inconsistent values. In some definitions editing includes also error correction. Imputation implies that missing or erroneous (e.g. edit failures) values for variables are replaced with imputed values, which have to be as correct as possible in regard to the true but unknown values. Imputation methods vary considerably depending on the type of data set, its scope and the type of missingness of data. In practice, editing and imputation are carried out in subsequent phases. Statistical editing is needed at each phase, starting from the planning of a data collection up to the formation of a data file, data processing and analysis. (Statistics Finland, 2007)

Editing model for Statistics Finland contains three main phases. The first phase of the model contains planning of editing process, descriptive analysis of data and error diagnostics. The second phase is the editing process in which the error identification and corrections are made. The last phase of the model evaluates the quality of both the editing and imputation process and the processed data. The editing model should be part of every statistical production process. By creating systematic process model for editing, Statistics Finland expects a clear improvement in efficiency of statistic process, but also improvement in quality and transparency of data. A big part of achieving the transparency of data for users and systematic quality control is to creating a list of indicator to be published.

In the Statistics Finland's editing model, actions for editing and imputation of statistical data are presented in process form. Indicators are involved in the editing model in two ways: On the one hand, they control error identification and correction actions and their effects on the data; on the other hand, they evaluate the development of the quality of data in different stages of editing process and overall quality of the final data.

We divide indicators for statistical data editing into three groups. In section 2 in this paper, we discuss indicators related to descriptive analysis of raw data. Indicators that relate to error identification are presented in section 3. Moreover, indicators that relate to the error correction are discussed in section 4. In section 5, we make an overview of the demands of the ESS standard of quality reporting in editing and outline recommendations for what indicators to use.

Indicators presented in this paper are collected from several sources: EUREDIT-project (EUREDIT / Ray Chambers, 2004), EDIMBUS-project (EDIMBUS, 2007), Eurostat standards for quality reports (EUROSTAT, 2009a), Quality Guidelines for Official Statistics by Statistics Finland (Statistics Finland, 2007) and functional report of BANFF-program by Statistics Canada (Statistics Canada, 2007).

2 Indicators of raw data

Indicators that describe raw data give information about errors on data, their effects on results and variables' or subgroups' significance on results. This initial editing and imputation is in the beginning of editing process. It may include observing the data in varying ways: tabulations, statistics calculations, distributional evaluation and data visualisation. Indicators describing the raw data are calculated in this phase.

Let *Y* be obtained data matrix (n x p), where n is number of observations and p is number of variables. For every variable y_j its observation y_{ij} *(i* = 1, ..., *n; j* = 1, ..., *p)* may have a value or it can be missing. The response indicator r_{ij} for observation y_{ij} has value of 1 if unit i has value on variable *j*. It has value of 0 when observation y_{ij} is empty. If value is missing due to structural reason, it should be marked. To identify structural missing values, we have factor b_{ij} that has value of 0 when y_{ij} is missing due to structural reason and 1 otherwise. Some indicators presented below are referring to auxiliary variable x. Variables x may originate from obtained data or from other sources. Survey weight for observation i is denoted by w_i.

First group of indicators we present include **indicators for missingness of data.** From two types of missingness (unit and item nonresponse) only item nonresponse is considered. Response rate (1), weighted response rate (2) for variable y_j and weighted response rate for variable y_j proportioned with auxiliary variable x (3) are defined in following table 1.

Table 1: Basic measures for missingness of data.

$\frac{\sum_i r_{ij}}{n}$ (1)	$\frac{\sum_i w_i r_{ij}}{\sum_i w_i}$ (2)	$\frac{\sum_i w_i x_i r_{ij}}{\sum_i w_i x_i}$ (3)

Weighted response rate (2) evaluates proportion of responses in variable *j* when whole population is inspected. This should be noted especially when data is stratified or quota sampling is used with deviant proportions or when calibration of weights are used. Weighted response rate for variable y_j proportioned with auxiliary variable x (3) gives information about effect of nonresponse to aggregates of variable y_j when x is correlated with y_j. Indicators for response rates, such as item nonresponse rate (4) and full response rate(5), are defined in table 2.

Table 2 : Indicators for response

$\frac{\sum_i(1-\prod_j r_{ij})}{n}$	(4)	$\frac{\sum_i(\prod_j r_{ij})}{n}$	(5)

Proportion of missing values in all variables (6) and average proportion of missing values (7) are presented in table 3.

Table 3: Indicators for measuring the proportion of missing values

$\frac{\sum_j(1-r_{ij})}{p}$	(6)	$\frac{\sum_i\sum_j(1-r_{ij})}{np}$	(7)

Proportion of missing values in all variables (6) is an unit-specific indicator of response. Proportion of missing values is usually reasonable to calculate in relation to some subset of variables $s < p$.

The effect of missing values on data can be evaluated with auxiliary variable x, which is available for all observations. In some cases it may be necessary to produce item nonresponse adjusted weights w_{ij}^* for variable j, where $\sum_i w_{ij}^* r_{ij} = \sum_i w_i$. Then it is possible to use indicators that evaluate the effect of missing values with auxiliary variable as presented in table 4.

Table 4 : Indicators evaluating the effect of missing values

$\frac{\sum_i w_{ij}^* r_{ij} x_i}{\sum_i w_i x_i}.$	(8)	$\frac{\sum_i w_{ij}^* r_{ij} x_i - \sum_i w_i x_i}{\sum_i w_i x_i}.$	(9)

Ratio of item nonresponse estimated and survey weight estimated totals of x (8) and proportion of variation of item nonresponse estimated and survey weight estimated totals of x (9) are defined above. If x and y are correlated, that ratio estimates the proportion of change that item nonresponse has on the total of variable y_j.

Next we will discuss on **indicators for impact of observation**. Variables that have skew distributions may have values, which have a large impact on results. Survey weights need also be taken into account when assessing the impact of observations. These indicators are more relevant when calculated from certain subgroup rather than from whole data, since the impact is easier to notice. It is also useful to specify essential subgroup by contents which significance on result can be calculated. Indicators for significance calculation are presented in table 5:

Table 5 : Significance indicators

$\frac{y_{ij}}{\sum_i y_{ij}}$ (10)	$\frac{\sum_{i\in q} y_{ij}}{\sum_i y_{ij}}$ (11)	$\frac{w_i}{\sum_i w_i}$ (12)	$\frac{w_i y_{ij}}{\sum_i w_i y_{ij}}$ (13)

Significance of each individual observation y_{ij} in sum of variable y_j (10) and similarly significance of each observation y_{ij} subgroup q (11) are typical indicators for significance examination. Significance of each weight w_i in sum of weights (12) identifies units that may have a large impact on results through survey weights. Significance of each weighted observation $w_i y_{ij}$ in estimate of total variable y_j (13) reveals the true impact of observation to total estimate. There are also some other indicators related to observation and weight significance. It is possible to calculate total estimate from subgroup that has one observation unit removed. This describes the effect the removed unit has on the estimate of total. These calculations require computing adjusted survey weights w_i^- with one unit removed accordingly.

Estimators sensitivity to unit *i*

$$c\big(\hat{\theta} - \hat{\theta}_{(i)}\big) \tag{14}$$

describes the change in estimate $\hat{\theta}$ when observation i is omitted. Term c standardizes the result and according to the EDIMBUS report, it can be formed as a mean of estimators on removed units *i:* $= \sum_i \hat{\theta}_{(i)}/n$. These estimates can be examined often through so-called sensitivity curve and it is linked with outlier evaluation.

Indicators presented above are mainly simple functions proportioned to total estimates. Significance evaluation has been crucial part of editing in recent years. In terms of selective editing there is several score functions available and most of them include reference value. These score functions will not be presented in this paper.

3 Indicators related to identifying of errors

Indicators concerning error recognition have two aims: 1) they describe the amount of errors in variables and observations, 2) they describe the effectiveness of error identification procedures. Indicators describing the efficiency of error detection are not discussed in this paper. Defects on data are evaluated in data studies and editing process phase on editing model. Error identification phase includes actions that are designed to identify errors so that they can be individualized on variables and observations.

The most common practise to notice error is to use an edit rule, which flags observation to be either error or error suspicion. Some error identification actions may include data processing with functions or modelling. Some errors are detected from results of macro editing. Visual examination on both observation and result levels is useful. It is essential for indicator calculations that identified errors and error rules used to identification can be tracked by flagged observations.

Error identification indicator f_{ij} for unit i on variable j has value of 1 if observation is detected to be erroneous on error identification process. Otherwise, it has value of 0. It is also possible to add denotation *l* for the parameter to identify method used on error detection f_{ijl}.

3.1 Indicators for error identification on different levels

Table 6 : Indicators for error identification: Variable level

$\frac{\sum_i f_{ij(l)}}{n}$	(15)	$\sum_i \sum_l f_{ijl}$	(16)

Indicators for error identification are needed in different levels of data. **On variable level**, error degree of variable y_j on data (by identification method *l*) (15), presented in table 6, measures the error identification rate of variable y_j. Adding the information of error identification method *l* describes the sensitivity of the method proportioned to all errors on variable. It is important to notice that error identification may include false alarms, so great number of identified errors may not imply a good error identification method. On the opposite, one sole error may be significant if the magnitude of error is exceptional. This is why it is sometimes useful to calculate the amount of error identifications on variable over all identification methods (16) for comparison. It also controls the operations of edit rules.

Indicator for error identification on observation level: Observation *i* is erroneous by edit rule *l* if at least one parameter f_{ij} $(j = 1, \ldots, p)$ has value of 1. Then error parameter e_{il} for observation i combined with edit rule *l* has value of 1 and otherwise value of 0. Variable amount p can be replaced with subset $(s < p)$ that are included on error identification. On the observation level of the data the proportion of error occurrence on all variables (by error identification method *l*)

$$\frac{\sum_j f_{ij(l)}}{p} \tag{17}$$

measures overall quality of observation.

Table 7 : Indicators for error identification: Data level

$\frac{\sum_i e_{il}}{n}$ (18)	$\frac{\sum_i(1-\prod_l(1-e_{ij}))}{n}$ (19)	$\frac{\sum_i w_i y_{ij} f_{ij}}{\sum_i w_i y_{ij}}$ (20)

Standard indicator for error identification **on data level,** presented in table 7, is error identification rate for edit *l* (18). If there are several edit rules in use it is possible to define general rate of error detection for all error detection methods (19). Proportion on error detection (20) describes the ratio of erroneous variables from total estimate.

4 Indicators related to correction actions

Indicators associated with error correction describe 1) quality of data after error correction; 2) amount of error correction in variables/units/data; 3) effect of error correction on results. After error identification and nonresponse assessment, we have gained information on to which observations and variables we should focus actions on. Actions might include inquiring the correct value from respondent, searching for right value from other sources, cold-deck imputation, imputation from other data source or imputation based on statistical methods. Let the inserted or imputed value be denoted as $\hat{y}_{ij}$. After different corrective actions we get the final data. It is quite common, that the edits on data have not been recorded in any ways and the only method to evaluate the changes is usually to compare the raw and final data.

Indicators for error correction can describe the final data and its flaws and edits. They remark on the broadness of edit actions, the amount of edits caused by error identification and influence of error identification to edit actions. In some cases, they show impact of faultiness on estimates. For calculating such indicators, it is essential to tag the information needed on observational level during the editing process.

4.1 Indicators measuring the proportion of missing values on data after error correction

The item response indicator $\hat{r}_{ij}$ takes the value of 1 if observation *i* on variable *j* has value *after E&I-actions*. The value can be the original y_{ij} or corrected $\hat{y}_{ij}$. Correspondingly $\hat{r}_{ij}$ takes value of 0 when y_{ij} is missing. Survey weight for observation *i* is denoted as w_i. Some standard indicators are presented in table 8.

Table 8 : Response and inconsistency rates

$\frac{\sum_i \hat{r}_{ij}}{n}$	(21)	$\frac{\sum_i w_i \hat{r}_{ij}}{\sum_i w_i}$	(22)	$\frac{\sum_i f_{ij} \hat{r}_{ij}}{n}$	(23)

The item response rate after E&I-actions (21) and weighted item response rate after E&I-actions (22) evaluates remaining missingness of variable j at population level. As with raw data, it is important also with final data to calculate weighted indicators if the data is stratified or quota sampling is used with deviant proportions or when calibration of weights is used. Usually it is recommended to calculate response rates with and without weights. The rate of inconsistent data (23) describes how much the variable j has values as a result of error detection focused on the variable j.

Table 9 : Observation specific indicators for missingness after imputation

$\frac{\sum_j (1 - \hat{r}_{ij})}{p}$	(24)	$\frac{\sum_i f_{ij} \hat{r}_{ij}}{p}$	(25)

As with the raw data, the proportion of missingness after imputation (24), presented in table 9, might be more informative to calculate in relation to some sensible subset s instead of all variables. Inconsistency proportion (25), presented in table 9, describes the portion of variables that has values as a result of error detection. The variable set of p can be substituted with subset of variables s that are included in error detection method.

Table 10 : Data specific indicators for missingness after imputation

$\frac{\sum_i (1 - \prod_j \hat{r}_i)}{n}$	(26)	$\frac{\sum_i (\prod_j \hat{r}_i)}{n}$	(27)	$\frac{\sum_i \sum_j (1 - \prod_j \hat{r}_{ij})}{np}$	(28)

Proportion of item nonresponse after imputation (26) and proportion of full response after imputation (27), presented in table 10, are exclusive classes, similarly to the situation with raw data, but both forms can be useful depending on situation. The remains of non-structural item nonresponse after correction actions indicate that it is not possible to form full response by current standards or data management system is enabled to allow some absence on observations. Mean proportion of missingness (28) describes how much missing values are included on observations on average.

4.2 Indicators describing error correction actions

Correction of an error is a modification of data and it is represented with variable $I(\hat{y}_{ij} \neq y_{ij})$, which takes value of 1 when value of raw data differs from value that is corrected and value of 0 otherwise. Parameter b_{ij} for structural missingness on raw data is included in some indicators and it's equivalent on corrected data is defined as $\hat{b}_{ij}$. It is highlighted in EDIMBUS-report that these indicators need to be calculated with and without weights. We start introducing **indicators for error correction actions on variable level** in table 11.

Table 11 : Indicators for error correction actions on variable level

$\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})}{n}$ (29)	$\frac{\sum_i w_i I(\hat{y}_{ij} \neq y_{ij})}{\sum_i w_i}.$ (30)	$\frac{\sum_i w_i I(\hat{y}_{ij} \neq y_{ij})\hat{y}_{ij}}{\sum_i w_i \hat{y}_{ij}}$ (31)

Edit rate (29) describes the quantity of edited values on specific variable. Also weighted edit rate (30) is defined. If we are only interested in imputed edits, terms are correspondingly imputation rate and weighted imputation rate. Edit ratio (31) describes the effect of edits proportioned to total estimate. Modification rate (32) is defined as

$$\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})(r_{ij}b_{ij})\hat{b}_{ij}}{n}. \tag{32}$$

Term $(r_{ij}b_{ij})\hat{b}_{ij}$ takes value of 1 only when variable has value given on observation$(r_{ij} = 1)$, it have not been structurally missing on raw data $(b_{ij} = 1)$ and it is not structurally missing on corrected data either $(\hat{b}_{ij} = 1)$. Similar indicators are net edit rate with term $(1 - r_{ij}b_{ij})\hat{b}_{ij}$, which takes account cases where response is formed for missing value, and cancellation rate with term $(r_{ij}b_{ij})(1 - \hat{b}_{ij})$, which takes account only removals of exiting values. These indicators can also be calculated with weights w_i.

If there is demand for indicators for specific correction method m, it is possible to calculate indicators above with a method specific parameter I_m as is presented in table 12.

Table 12 : Edit proportions and overall edit rate

$\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})I_m}{\sum_i I(\hat{y}_{ij} \neq y_{ij})}$ (33)	$\frac{\sum_j I(\hat{y}_{ij} \neq y_{ij})}{p}.$ (34)	$\frac{1}{p}\sum_j \left(\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})}{n}\right).$ (35)

Proportion of correction method m from all correction methods used for variable y_j (33) can also be calculated with the weights w_i. Indicators presented above have also their corresponding versions in observation level, describing the change within the observation from different starting points, for example edit proportion in observation level (34). Modification proportion, net edit proportion and cancellation proportion can be calculated by using terms $(r_{ij}b_{ij})\hat{b}_{ij}$, $(1 - r_{ij}b_{ij})\hat{b}_{ij}$ and $(r_{ij}b_{ij})(1 - \hat{b}_{ij})$ correspondingly. Set of variables p can also be substituted with variable subset s. There are also similar indicators for describing edits on data level, for example overall edit rate (35) presented in table 12. Overall modification rate, overall net edit rate and overall cancellation rate can be calculated correspondingly. In addition, weighted versions are available by adding weight parameter.

4.3 Indicators for implications of error identification

Not every error identification results in an error correction. On next table 13, we present some standard indicators for implications of error detection.

Table 13 : Indicators for implications of error detection

$\frac{\sum_i f_{ij} I(\hat{y}_{ij} \neq y_{ij})}{\sum_i f_{ij}}$ (36)	$\frac{\sum_i e_{ij}\, I(\hat{y}_{ij} \neq y_{ij} \mid l)}{\sum_i e_{ij}}$ (37)	$\frac{\sum_i w_i(\hat{y}_{ij} - y_{ij})}{\sum_i w_i}$ (38)	$\frac{\sum_i w_i(\hat{y}_{ij} - y_{ij})}{\sum_i w_i\, y_{ij}}$ (39)

It is possible to calculate edit rate proportioned to errors detected for variable *j* (36). It describes amount of corrections resulted from error detection proportioned to all errors included in variable. Such edit rate is also possible to define for specific detecting method *l* respectively, assuming that there is certainty that an edit resulted only from inspected method. For this reason, it is necessary to define conditional indicator $I(\hat{y}_{ij} \neq y_{ij} | l)$, which takes value of 1 only when value modification has resulted from error detection method *l*. Now we can define rate for error corrections in variable *j* caused by error detection method *l* (37) and it is possible to extend for data level.

There are also defined indicators that describe the impact of corrections and error quantity on data. Error corrections have effect on results and this effect can be evaluated with indicators that are based on value difference $\hat{y}_{ij} - y_{ij}$ or estimate difference $\hat{\theta}(\hat{y}_1, \ldots, \hat{y}_i, \ldots, \hat{y}_n) - \hat{\theta}(y_1, \ldots, y_i, \ldots, y_n)$. If there is information on real values y_{ij}^* it is possible to evaluate defects of error correction with similar indicators by replacing the raw data values y_{ij} with real values y_{ij}^*. Some common variable level indicators for measuring edit impact are weighted average edit impact (38) and weighted relative average edit impact (39) presented in table 12. Weighted α-relative edit impact

$$\frac{\left(\frac{\sum_i w_i (\hat{y}_{ij} - y_{ij})}{\sum_i w_i y_{ij}}\right)^{1/\alpha}}{\sum_i w_i y_{ij} / \sum_i w_i} \tag{40}$$

is average correction impact modified with α proportioned to variables weighted mean $\bar{y} = \frac{\sum_i w_i y_{ij}}{\sum_i w_i}$. With α it is possible to regulate examination and when $\alpha = 1$ indicator results in weighted average impact of edits proportioned by variable *j*. Same examination can be done for correction defects.

Total impact of edits can be calculated as difference of estimates

$$\hat{\theta}(\hat{y}_1, \ldots, \hat{y}_i, \ldots, \hat{y}_n) - \hat{\theta}(y_1, \ldots, y_i, \ldots, y_n)\,. \tag{41}$$

Sensitivity of the parameter estimate $\hat{\theta}$ for edits is denoted as

$$c\left(\hat{\theta}(\hat{y}_1, \ldots, \hat{y}_i, \ldots, \hat{y}_n) - \hat{\theta}(\hat{y}_1, \ldots, \hat{y}_{i-1}, y_i^*, \hat{y}_{i+1}, \ldots, \hat{y}_n)\right) \tag{42}$$

where *c* is a suitable standardization constant (usually *c* = 1). It describes the change on estimate that is based on corrected values when one corrected value is replaced by real value. Edit error rate

$$\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij}) I(\hat{y}_{ij} \neq y_{ij}^*)}{\sum_i I(\hat{y}_{ij} \neq y_{ij})} \tag{43}$$

describes amount of false edits proportioned to all edits.

If all edits under examination are done by imputation, we can refer to indicators above as *weighted average imputation impact* (38), *weighted relative average imputation impact* (39) etc.

5 Recommendations for use of indicators

Quality standards of Eurostat: Eurostat has determined standards for quality reports for European Statistical System. The guidebooks of standards describe different dimensions of quality and in this paper, we will refer to the following demensions: measurement errors, nonresponse errors, processing errors in micro-data and imputation.

A measurement error is the discrepancy between the observed value of a variable provided by the survey respondent and its underlying true value. Eurostat reports discuss the reasons for measurement errors, how they are formed and what methods exists to investigate measurement errors. Error identification offers information on certain and suspected errors, and indicator for *error identification rate for edit rule l* (18) is mentioned. It is also recommended to use this indicator to examine crucial subgroups. For bias caused by measurement errors, they offer different evaluation studies that are mainly related to assessing the questionnaire and the interview situation. One way to estimate bias is to calculate results from original data and final data and compare them.

Item nonresponse on variable is crucial of nonresponse errors according to Eurostat report. Indicator for *item response rate* (1) is mentioned and optional indicator for *weighted item response rate* (2) is offered. It is important to specify the essential variables for which the response rates are calculated and consideration must be used to decide whether to use weighted or unweighted indicators. Examination in relevant subsets of units is also highly recommended. The effect of nonresponse can be tested with values that are available for all responded units by comparing full response estimate to the estimate that notices nonresponse. Indicators for nonresponse estimates are mentioned for example *Ratio of item nonresponse estimated and survey weight estimated totals of* x (8).

Processing errors in micro-data imply on data recording, editing and sometimes on data treatment and imputation. By Eurostat report, it is essential to explain the extent and impact of processing errors if they are significant. Calculating results from original and corrected data and comparing the results is mentioned as a simple testing method. This provides the total net effect of editing. The amount of imputation, or generally error correction, can be evaluated by calculating *edit rate* (35).

Recommended and considered indicators for different purposes: Concepts used in indicator calculations are not always unambiguous. In next section we will discuss on different situations and problems related to these concepts. ***Target variables:*** Statistic producer must decide the group of variables from which the indicators are calculated. Not all indicators can be calculated for categorical variables. Some variables are too insignificant for detailed examination. Therefore, it is essential to define certain target variables of which indicators are calculated. ***Subgroups:*** It is possible that indicators perform better on subgroup level than in complete data. Knowledge of contents and overall experience are needed to define such subgroups. ***Raw data:*** There are several indicators targeted for raw data examinations, but it is not always obvious what is considered as raw data. For the basis of the editing model, raw data is defined as data that includes all needed material combined. It is possible that initial error check has been done to some parts of data when it has been received and before the raw data has been formed. In some statistics editing is greatly emphasized on data reception phase. All these initial edit actions must be regarded when indicators from raw data are calculated.

Structurally missing values are problematic with indicators that describe item nonresponse. In some cases, this can be taken into account with separate parameter. If structurally missing values exists in raw data, they must be defined and their effect on indicators should be eliminated. ***Error identification:*** Different methods for error identification are categorized in different phases in editing model. For all cases, it is not possible to define exactly the method of error identification due to unsystematic detection. There are indicators that are used to identify error detection methods but they are mainly designated for edit rules. Sometimes error identification method is not necessarily important to notice. Crucial identification methods should be selected based on criteria of the editing model and indicators should be targeted on those identification methods. In multiple error cases, it is not apparent whether to tag all errors on the value or just the error that caused disqualification.

Error correction: Main problem with error corrections is what corrections are included in indicator calculations. Original value can be changed unambiguously because of logical inspection or inquiry of the correct value. These cases might not be necessarily involved in indicator calculations. When error correction indicators are concerned usually only imputation is mentioned. Generally, only significant error correction methods are included in indicator calculations. There might be problems with individualizing correction methods and with serial corrections. Some balance adjusting edits also modify values that are not erroneous.

6 Discussion

In this paper we have collected indicator in respect to editing process of official statistics. We have classified these indicators in three groups according to their functions: Indicators of raw data, indicators that relate to the error identification and indicators that relate to error correction. These groups relate to three different phases of editing model. The number of indicators presented in this paper is quite substantial considering how diverse statistics production processes are. Not all indicators are suitable for every type of statistics. Hence, consideration on which indicators to be applied in each process is essential. Many questions related to choices of subgroups or variables from which indicators are calculated need solid substance knowledge. Some indicators presented are important but suitable only for specific situations. There for, it is not possible to define a detailed list of indicators to be published with all statistics. Some standard indicators for editing process should always be computed. Indicators measuring missingness in data (presented in section 2.1) are valuable tools for statistics production process by describing the coverage of data in each stage of process. Then, if any editing actions are done, the user of the final product should have access to information on edit rates of the data (presented in section 4.2). Altogether, indicators are essential part of editing process as they provide information on quality of data, results and editing process in general.

References

EDIMBUS (2007). Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys. Project report. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

EUREDIT / Ray Chambers (2004). Evaluation Criteria for Statistical Editing and Imputation, Project report.

EUROSTAT (2009a). ESS Standard for Quality Reports, Luxemburg.
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/ESQR_FINAL.pdf

EUROSTAT (2009b). ESS Handbook for Quality Reports, Luxemburg.
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/EHQR_FINAL.pdf

OLLILA, P. (2012). Raaka-aineiston,editoinnin ja imputoinnin indikaattorit (luonnos). Unpublished memorandum. Statistics Finland.

OLLILA, P. and ROUHUVIRTA, H (2012). Editoinnin prosessimalli (luonnos). Unpublished memorandum. Statistics Finland.

STATISTICS CANADA (2007). Functional Description of the Banff System for Edit and Imputation, Ottawa.

STATISTICS FINLAND (2007). Quality Guidelines for Official Statistics 2nd Revised Edition, Helsinki.
http://www.tilastokeskus.fi/meta/qg_2ed_en.pdf

Evaluation labour input of filling in statistical forms: sampling methods

Julia Orlova[1]

[1]Belarus State Economics University, Minsk, e-mail: orlova-julia-gen@mail.ru

Abstract

The article substantiates the need for the regular evaluation of the labour input of respondents filling forms of statistical reporting. The intensity measurement approaches are shown, the use of selective monitoring in this area is explored. It also considers the problems of defining the scope and the design of sample survey of labour input associated with filling the forms of statistical reporting.

Key words: labour input, statistical forms, sampling.

1 Introduction

The increasing information needs of users in official statistics requires updating and improvement of existing forms of State statistical surveys and instructions for their filling, approving of new forms and relevant indicators and cancel obsolete ones. Particular attention in the revision of State statistical reporting in Belarus will be given to comments and suggestions of the reporting institutions (respondents) in order to minimize labour input associated with filling the forms of statistical reporting.

In the process of measuring labour input of statistical work, in particular the works of respondents, a range of questions is usually raised: the adequate estimation of elapsed time, its structure, the lack of normative work and as a result, there is difficulty establishing clear boundaries research facility; the absence of the similar survey in national statistical practice, discussions of the use of continuous monitoring techniques (there are 112 statistical reporting forms and 126771 legal entities on June 1, 2012).

This paper discusses the research and the objectives of the labour input studying. Taking into account the available information of State statistics authorities the direction of basics, volume and sampling design is proposed.

A respondents' sample survey is proposed to be the main source of information on the labour input connected with filling State statistical reporting. Until now, the sample survey of labour input associated with filling the forms of State statistical reporting in Belarus has never been carried out. In 2012, a survey has been carried out for the purpose of obtaining empirical data about the labour input of filling State statistical reporting, total and by economic activities of responders. Survey object is legal entities of the Republic of Belarus.

2 Evaluation labour input of filling in statistical forms

In order to form the general population of the respondents they were proposed to fill in such a questionnaire (table 1).

Table 1: Questionnaire study of the response burden associated with completion of forms of state statistical reporting

The operative time costs, hours					
up to 1	1-2	2-4	4-8	8-40	more than 40

In assessing the complexity of time spending on filling in statistical forms it is proposed to distinguish the following categories of the working time: operative time, set-up time, while observing the operation of the equipment. At the same time the operative work must be submitted to time, which is directly aimed at the implementation of the set tasks.

Thus, the operative time includes the time required for registration and reporting forms of state statistical reporting; primary input records to the PC, the union of sets of information received from affiliates or divisions, the checks on the statistical reporting forms: arithmetic, logical, comparison with the previous period and so on, forming the output tables, etc.

The set-up time includes time to prepare for the implementation of a given work and time to perform the operations associated with its ending.

The set-up time includes consultations in the area (businesses) on issues arising in the process of monitoring, the workers' press reports and organization of the materials for controlling details of statistical reporting.

While observing the operation of the equipment there was included the time to perform operations such as monitoring the operation of the equipment when working with input and output information.

On the basis of survey conducted in government statistics and showed in this study, the author proposes the following system of correction factors developed by time consuming on filling in statistical forms:

1. complexity (Kc);
2. occupancy (Ko).

The average occupancy rate for each form of statistical reporting is a known quantity. Thus, as the result of the research we receive the complexity rate for each of the statistical reporting forms.

3 Evaluation labour input of filling in statistical forms: sampling methods

3.1 General

To organize specialized sample surveys of labour input associated with filling the statistical forms the general population of the responders must be formed on grounds of the operative time costs at the level of the Republic (area, region). At present, the general population can be formed on the basis of lump-sum continuous questioning conducted by the State statistical bodies in 2012, the questionnaires (table 1) were attached to each statistical form received by legal entities.

3.2 Sample basis

A survey carried out by statistics bodies has allowed forming database, which is used to select the sample units. The sampling frame is formed by legal entities and their separate units, which should be grouped on a territorial basis and on the basis of economic activity.

Table 2: Distribution of legal entities by regions and in Minsk on June 1, 2012

Belarus	Total	Percentage
	126 771	100,0
Including regions:		
Brest	13 762	10,9
Vitebsk	13 085	10,3
Gomel	13 324	10,5
Grodno	11 042	8,7
Minsk city	43 844	34,6
Minsk	20 361	16,1
Mogilev	11 353	9,0

Table 3: Distribution of legal entities by economic activity on June 1, 2012

Belarus	Total	Percentage
	126 771	100,0
agriculture, hunting and forestry	4 731	3,7
fishing, fish farming	212	0,2
mining industry	75	0,1
manufacturing	15 826	12,5
production and distribution of electricity, gas and water	228	0,2
construction	10 071	7,9
trade; repair of motor vehicles, motorcycles and personal and household goods	42 793	33,8
hotels and restaurants	2 487	2,0
transport and communications	8 948	7,1
finance operations	558	0,4
operations with real estate, renting and services to consumers	15 952	12,6
government	4 392	3,5
education	8 788	6,9
health and social services	2 074	1,6
provision of utilities and other services	9 636	7,6

On the basis of the limiting errors calculation, depending on the sample size and values of characteristics (with 95% probability level), the recommended sample size is 22500 entities that makes up 25% of the general population (table 4).

Table 4: Limiting errors in dependence on sampling size

Sample size	Limiting error, %	Limiting error, units
5%	1,13	1 688
10%	0,77	1 162
15%	0,61	922
25%	0,45	671
30%	0,39	592

3.4 Sample design

While evaluating the labour input of filling in statistical reports the general population of Belarus legal entities should be stratified by territorial characteristics of the responders to receive representative estimates at the regional and national levels. The general population of legal entities should be also stratified by the identity of the respondent to the form of economic activity.

References

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

Sharon L. Lohr's. Sampling: Design and Analysis. 2010.

Real donor imputation pools

Nicklas Pettersson[1]

[1]Stockholm University, e-mail: nicklas.pettersson@stat.su.se

Abstract

Real donor matching is associated with hot deck imputation. Auxiliary variables are used to match (donee) units with missing values to a set of (donor) units with observed values, and the donee missing values are 'replaced' by copies of the donor values, as to create completely filled in datasets. The matching of donees and donors is complicated by the fact that observed sample survey data is both sparse and bounded. The important choice of how many possible donors to choose from involves a trade-off between bias and variance. We transfer concepts from kernel estimators to real donor imputation. In a simulation study we show how bias, variance and the estimated variance of a population behaves, focusing on the size of donor pools.

Keywords: Bayesian Bootstrap, Boundary and nonreponse bias; Multiple imputation

1 Introduction

Missing data is always a nuisance. The 'holes' in the dataset precludes many simple standard techniques. Datasets obtained by excluding partially observed units (e.g. due to item nonresponse) give inefficient and usually quite biased results. A better alternative is to impute the missing values. An extensive book on missing data state that (p72, Little & Rubin, 2002)

"Imputations should generally be:
(a) Conditional on observed variables, to reduce bias due to nonresponse, improve precision, and preserve association between missing and observed variables;
(b) Multivariate, to preserve associations between missing variables;
(c) Draws from predictive distribution rather than means, to provide valid estimates of a wide range of estimands.".

The most important factor in imputation is access to auxiliary variables which are predictive of the missing values and the nonresponse propensity. Real donor (hot deck) imputation (Laaksonen, 2000) uses auxiliaries to match a donee unit with missing values to a set (pool) of close (nearest neighbour) donor units with observed values, and then 'replaces' the donee missing values by copies of randomly drawn donor values. It is often applied within cells from cross-classified categorical (and sometimes subjectively classified continuous) auxiliares. We only discuss continuous variables with univariate missingness. Point (b) is therefore not relevant here.

Point (c) relates to multiple imputation (Little & Rubin, 2002), which is a method for representing missing data uncertainty. The missing values are then imputed several times, and each imputed dataset is analyzed separately. The final estimates consists of the pooled results.

The size of donor pools becomes important here. Pools with few potential donors give rise to strong correlation between the values imputed for a missing value. In repeated sampling this results in highly variable final estimates, similar to sampling from correlated (e.g. clustered) data. Larger donor pools may instead reduce the quality of matches and increase the bias. The number of potential nearest neighbours donors thus regulates the trade-off between bias and variance in imputation, in parallell with pointwise kernel estimators. Features from this area have been applied in imputation to deal with the sparse and bounded data (Aerts et al, 2002; Pettersson, 2012), and to decide the donors pools (Schenker & Taylor, 1995; Marella, Scanu and Conti, 2008). We discuss these issues in the following sections. In simulations we show how different strategies for selecting the number of donors and the features for bias reduction from Pettersson (2012) affects bias, variance, and estimates of variance of a population mean estimate. To yield valid inference, our method is based on the Bayesian Bootstrap (Rubin, 1981).

2 Selecting the donor pool

The choice of bandwidth is important in kernel estimation. Several types of bandwidths exists. A fixed bandwidth corresponds to having imputation donor pools consisting of units with a (auxiliary based) distance to the donee which is less then a value ϵ. A fixed 'rule-of-thumb' bandwidth based on distributional assumptions is often a good starting point (Silverman, 1986). Fixed bandwidths can be locally adapted by increasing (decreasing) the maximum allowed distance if relatively few (many) donors are close to the donee, i.e. if the density at the donee auxiliary value is low (high). Always using the same number of potential donors in all donor pools corresponds to a nearest neighbour bandwidth, which may find donors that are better matched to the donee in densely regions, and automatically ensures that no donee get zero donors. The trade-off between bias and variance means that gains in precision from increasing the number of donors may result in reduced quality of the matches and increased bias. Different estimators may profit from different strategies of choosing the donor pool size/bandwidth.

3 Bias reduction

A disadvantage of the real donors' methods is that a donee and its pool of donors usually are imperfectly matched. Particularly, this becomes a problem when the donee auxiliary values lies at the boundary (i.e. convex hull) of the donors auxiliary values, since there may be no or only a few potential donors with observed auxiliary values that lies on one side of the donee auxiliary value. The donor pool is then badly balanced to the donee. If such a pool is used for imputation, the risk is also larger that bias is introduced in the imputed study variable.

Pettersson (2012) employs three methods to reduce this bias. First, since the closest donors provide a better match to the donee, they are given higher selection probabilities than more distant donors. Due to the optimality properties in estimation the donor selection probabilities are decided by an 'Epanechnikov' function (Silverman, 1986). Secondly the selection probabilities are calibrated so that the expected imputed auxiliary value equals the auxiliary value of the donee. The third method not only reduces the selection probabilities but also completely removes the furthest donors in the pool (which matches the donee least and thus contributes most bias), and only keep the best matches which gain larger selection probabilities. The bias will be reduced, but donor pool variance is expected to increase.

4 Simulation

We used the setup in Pettersson (2012) with a population of $N = 1600$ units, from which $G = 1000$ samples of size $n = 400$ was drawn, and with each study variable imputed $B = 20$ times using the auxiliary variable from which it was generated. Since bandwidth behaviour may depend on the underlying distribution we used three auxiliaries; $X_{Uniform} \sim U(\pi/6, 2\pi)$; $X_{Normal} \sim N(13\pi/12, 11\pi/48)$; and $X_{Gamma} \sim Z + \pi/6$, where $Z \sim Gamma(1, 1/2)$. All auxiliaries approximately had a range of $(\pi/2, 2\pi)$, where X_{Gamma} had an outlier at 6.28. We choose a logistic missingness mechanism $logit(Pr(R = 1|X) = -1 + \beta_z \sum_{i=1}^{5}(X - (2i\pi - 2)/4)^2$, where β_z were adjusted to give on average 25% missingness irrespective of the auxiliary ($z = Uniform, Normal, Gamma$).

Imputation methods relies on the relation between study and auxiliary variables, so we generated a linear $Y_X = X_z + e_{X_z}$, a nonlinear $Y_{cosX} = cos(4X_z) + e_{cos(4X_z)}$, and a mixed $Y_{X+cosX} = X_z + cos(4X_z) + e_{X_z+cos(4X_z)}$ study variable. The error terms e_t were generated from $N(0, Var(t))$. The probability of nonresponse on the study variables induced by the missingness mechanism is thus highest (lowest) as $cos(4X_z) = 1(0)$. Means and (co)variances are found in table 1.

The number of potential donors k was determined in three ways. The first method (knn) initially used $k = 2, ..., 30$ potential donors, and gradually increased the number as more values were imputed. Secondly, we used a rule-of-thumb method (fix) where the donor pool consisted of units with distance less than $\epsilon \propto s_{X_z} m^{-1/5}$ from the donee, where m is the number of potential donors. Thirdly, we used a locally adapted version ($adap$) of fix, where ϵ was increased (decreased) if the density at the donee auxiliary value was low (high) (see Silverman p101, 1986). We also used versions with the bias reduction features from section 3 added, (knn_b, fix_b and $adap_b$).

We compute; $Bias = \frac{1}{G}\sum_{g=1}^{G}(\widehat{\overline{Y}}_g - \overline{Y})$, where $\widehat{\overline{Y}}_g = \sum_{b=1}^{B}\widehat{\overline{Y}}_{b,g}$ is the overall estimated mean in the B imputed datasets and $\overline{Y}$ is the true mean; $Var = \frac{1}{G}\sum_{g=1}^{G}(\widehat{\overline{Y}}_g - \frac{1}{G}\sum_{g=1}^{G}\widehat{\overline{Y}}_g)^2$; and relatve error of estimated variance $\frac{Var-\widehat{Var}}{Var}$, where $\widehat{Var} = \frac{1}{G}\sum_{g=1}^{G}(s_{Y_g} + \frac{B+1}{B(B-1)}\sum_{b=1}^{B}(\widehat{\overline{Y}}_g - \widehat{\overline{Y}}_{b,g})^2)$ is the average estimated variance.

Table 1: Means and (co)variances of simulated data

	Mean			Variance			Covariance with X_z		
z	u	n	g	u	n	g	u	n	g
Y_{X_z}	1.08	3.47	3.42	3.04	3.19	5.65	0.32	0.19	0.51
Y_{cX_z}	-0.44	0.04	0.01	0.62	0.76	0.72	0.51	-0.04	0.48
$Y_{X_z+cX_z}$	0.64	3.51	3.44	6.08	5.95	8.74	2.89	0.17	3.08
X_z	1.03	3.42	3.38	0.27	0.51	2.81	0.27	0.51	2.81

5 Results

We present the results in figures 1-3. The number of initial donors for knn and knn_b is plotted against bias, variance and the relative error in estimated variance. We also add horizontal lines for fix, fix_b, $adap$ and $adap_b$. Due to the shrinkage feature, the initial number of donors is expected to be larger then the final number of donors for the bias corrected methods.

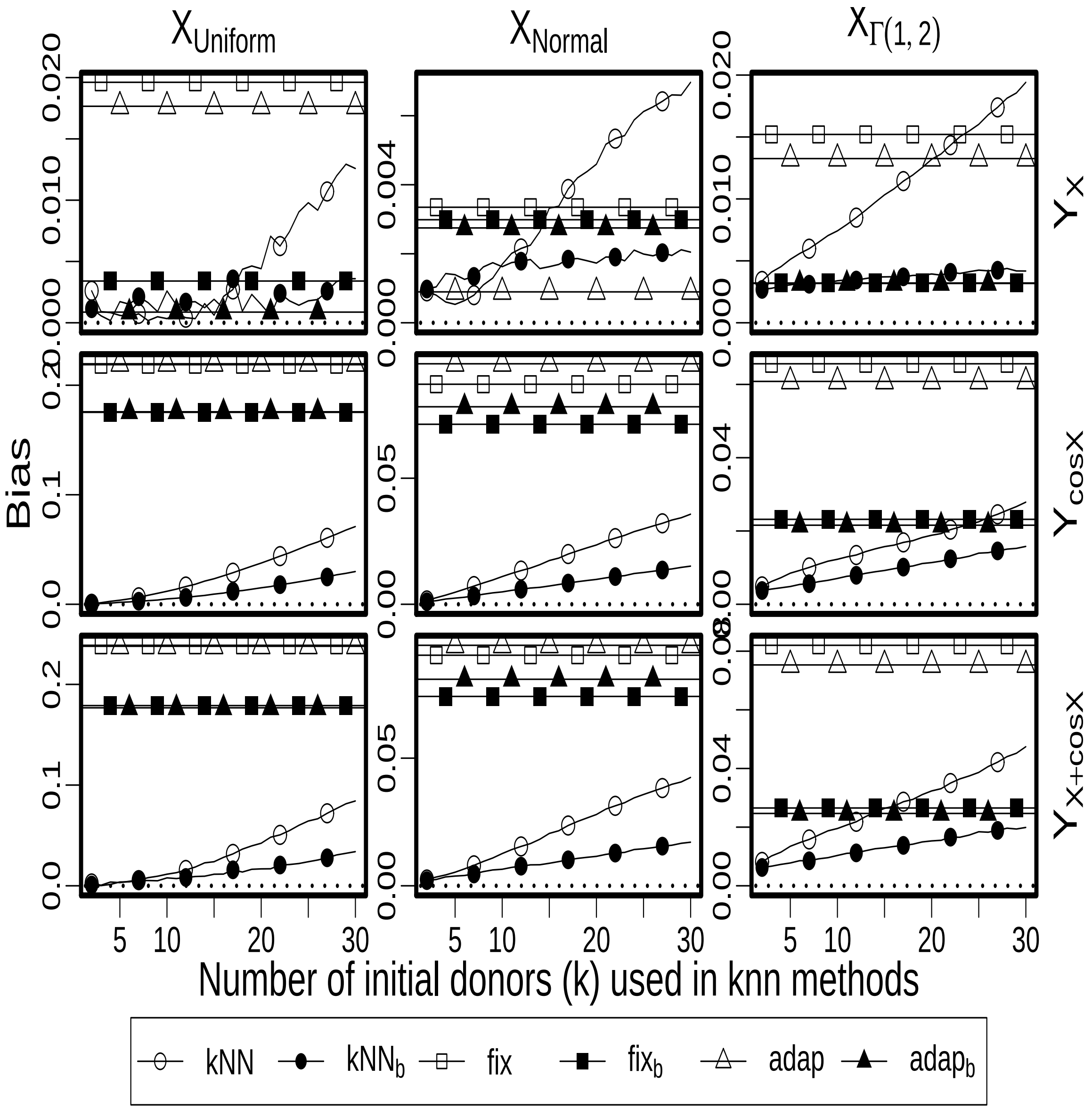


Figure 1: Bias of estimates from simulations

Except for the least complex data $Y_{X_{uniform}}$ and $Y_{X_{normal}}$ with small initial k, bias is always smaller for knn_b compared to knn. Bias tend to increase as k increases for both methods, but knn_b at a lower rate. knn_b also has lower bias than the fixed and adaptive versions with a few exceptions for Y_{X_z} when they are comparable. Bias corrected versions fix_b and $adap_b$ always give lower bias than its noncorrected counterparts fix and $adap$, except for $Y_{X_{normal}}$ with $adap$.

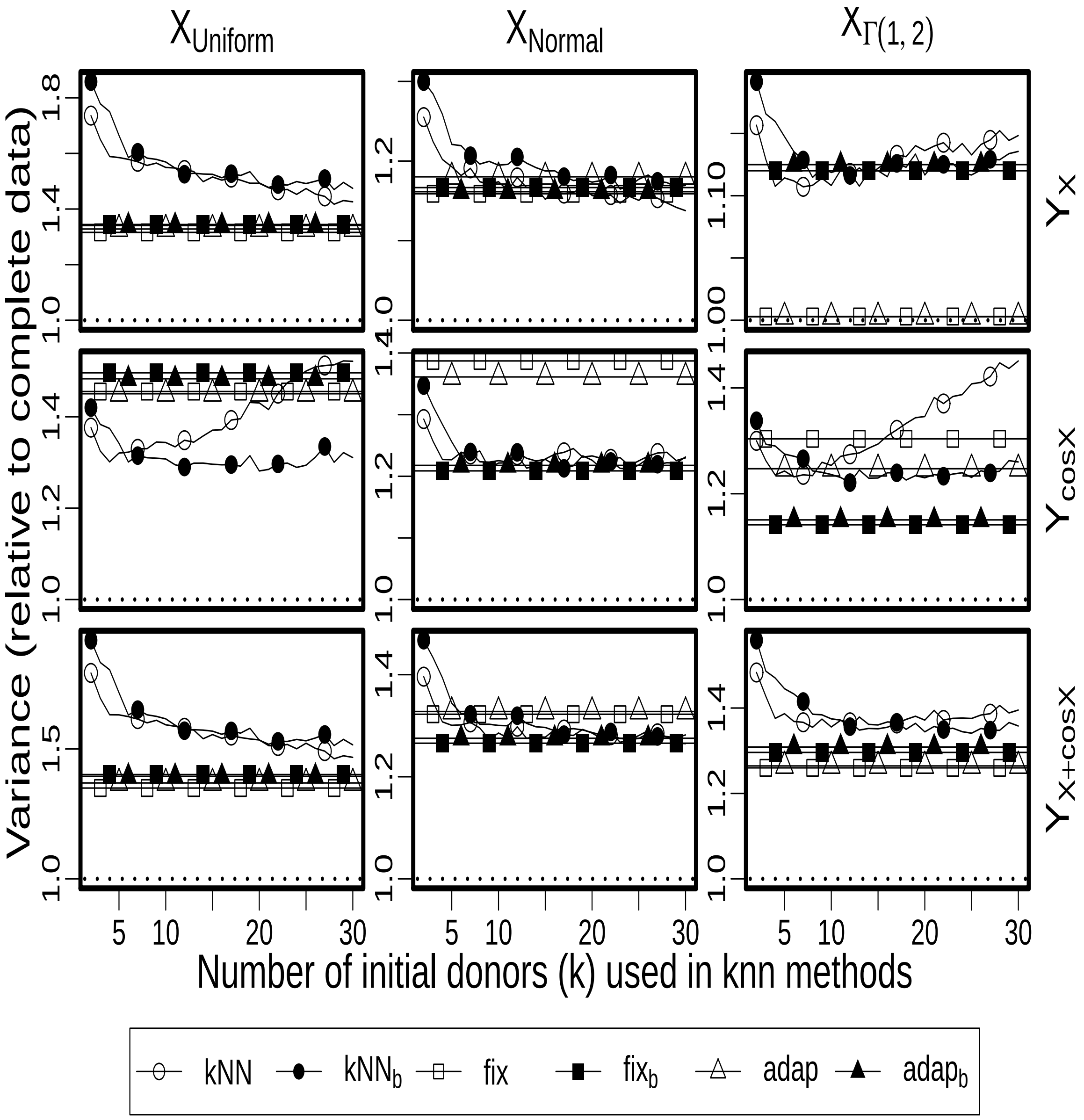


Figure 2: Variance of estimates from simulations

For small k, variance always falls as k is increased but is slightly higher for knn_b compared to knn. For larger k variance continue to fall or flatten out, except for knn where it sometimes increase, especially with auxiliary X_{Gamma}. Both fixed and adaptive methods generally have lowest variance, but nearest neighbours usually approached them as k was increased, and knn_b was always lower for $Y_{cosX_{uniform}}$. For $Y_{X_{uniform}}$ methods without bias correction (which on average also used most donors) had variance not far from complete data.

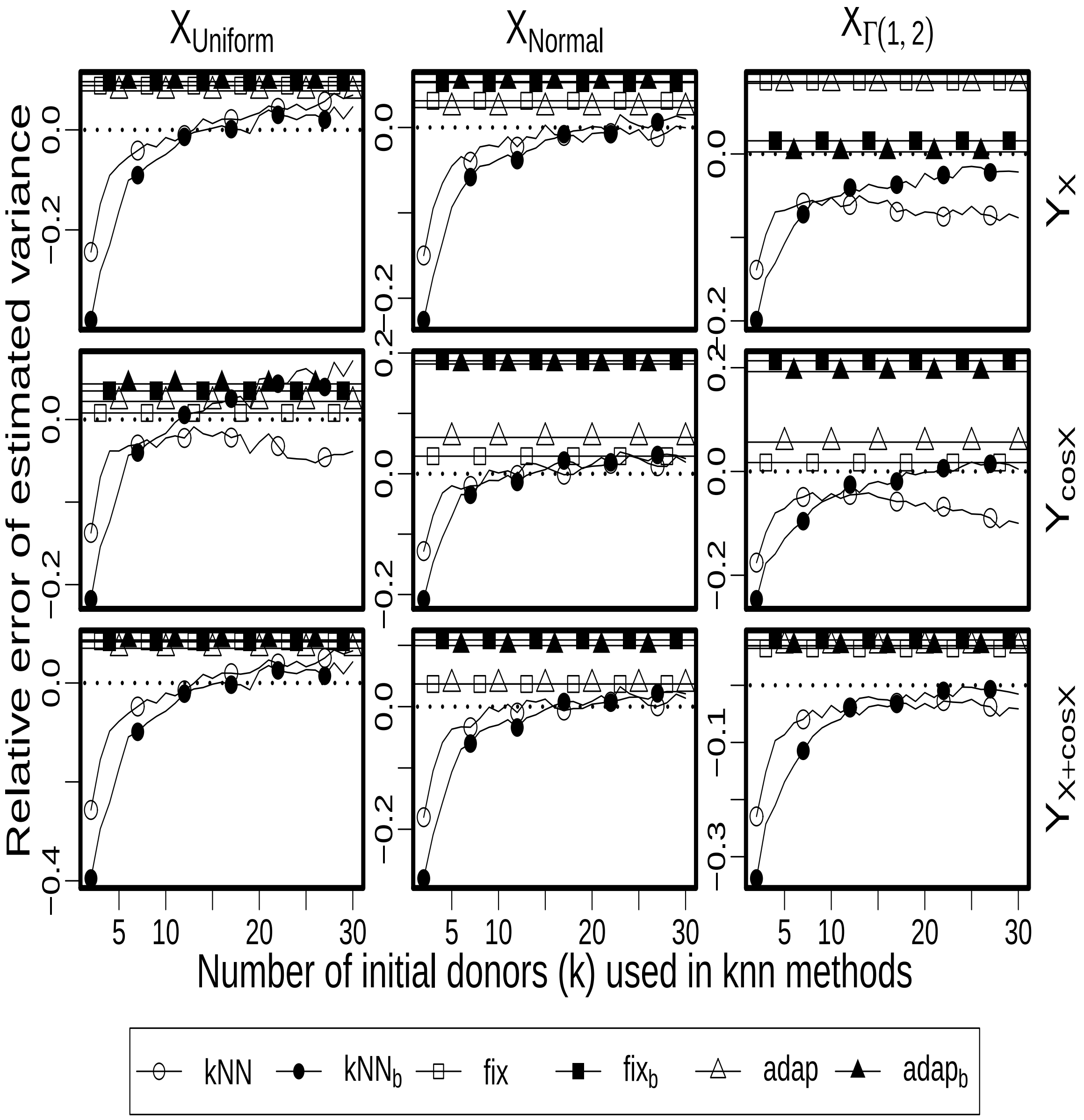


Figure 3: Relative error of estimated variance of estimates from simulations

6 Conclusions

Multiple real donor imputation has the advantage of requiring few model assumptions and imputing observed values. But some difficulties with continuous auxiliaries arise that needs to be dealt with. Since boundary donee units with missing values can only be matched to donors on one side, donor pools will be biased. Relative sparseness of donors also worsen the probability of forming good predictive donor pools. The size of donor pools is important since it involves a trade-off between bias and variance and affects the ability of estimating variances. This was clearly seen here where the fixed/adaptive methods, which generally had large donor pools, also had larger bias but smaller variance. Without any bias reduction applied there is certainly a risk of increased bias (and variance) associated with increasing donor pool sizes. Increasing the number of donors for boundary donees naturally worsens the already insufficient matching. Too few donors is on the other hand associated with high variance and too low variance estimates. The bias reduction techniques adresses the boundary bias and matching by adapting the donor pools. Given sufficiently many initial donors, it can make bias of the nearest neighbour method less dependent on the exact number of donors, and also improve bias of fixed/adaptive methods. We only study one fixed (and adaptive) rule-of-thumb method in our simulation, and blind use of it obviously involved a risk of getting large bias. Compared to a reasonably large nearest neighbour it only hade lower MSE when the study variable was a linear function of a uniform auxiliary. This seems to be associated with its generally larger donor pools giving rise to larger bias. Simulations with several other fixed methods (not presented here) generally also gave larger bias but smaller variance then nearest neighbour methods. The effects from local adaption of the fixed method seemed relatively small here and need further investigation.

References

Aerts, M. Claeskens, G. Hens, N. Molenberghs G., (2002). Local multiple imputation. *Biometrika*, 89(2), pp.375-388.

Laaksonen, S., (2000). Regression-based nearest neighbour hot decking, *Computational Statistics*, 15(1), pp.65-71.

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

Marella, D. Scanu, M. Conti, P.L., (2008). On the matching noise of some nonparametric imputation. *Statistics and Probability Letters*, 78, pp.1593-1600.

Pettersson, N., (2012) Bias reduction of finite population imputation by kernel methods. *To appear in Statistics in Transitions.*

Rubin, D.B., (1981). The Bayesian bootstrap, *Annals of Statistics*, 9, pp.130-134.

Schenker, N. Taylor, J.M.G., (1996), Partially Parametric Techniques for Multiple Imputation, *Computational Statistics and Data Analysis*, 22, pp.425-446.

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

Inclusion probabilities for successive sampling

Tomas Rudys[1]

[1]Statistics Lithuania, e-mail: tomas.rudys@gmail.com

Abstract

We give a short overview of calculation of first and second-order inclusion probabilities for successive sampling design. We compare the successive sampling first and second-order inclusion probabilities with already known numerical approximation results for Pareto πps and Conditional Poisson sampling designs. Pareto πps and successive sampling was introduced by Rosén and belongs to a class of sampling designs called order sampling with fixed distribution shape. At first an order sampling is introduced. We also give the examples of calculation of first and second-order inclusion probabilities for the mentioned above different sampling designs.

Keywords: Order sampling, successive sampling, first and second-order inclusion probabilities, numerical integration.

1 Introduction

Rosén (Rosén, 1996) studied and introduced a class of sampling designs called order sampling designs, which are executed as follows. Independent random variables, called ordering variables, are associated with the units in the population. A sample of size n is generated by first realizing the ordering variables, and then letting the units with the n smallest ordering variable values constitute the sample. Rosén also defined order sampling designs with fixed distribution shape: uniform, exponential, Pareto, successive. Author also derived the exact formulas for calculation of inclusion probabilities (Rosén, 1998) for these sampling designs.

Krapavickaitė (Krapavickaitė, 2012a) showed that Lithuanian Labour Force survey has successive sampling design and analysed the quality implementation actions for Lithuanian Labour Force Survey. Krapavickaitė also analysed order sampling designs and gave formulas for calculation of first and second-order inclusion probabilities for successive sampling (Krapavickaitė, 2012b).

Conditional Poisson and Pareto πps sampling designs were analysed and compared by Aires (1999) where the algorithms to find exact inclusion probabilities were derived. Author showed that it is feasible to calculate first and second-order inclusion probabilities for both sampling designs and program routines provide good numerical precision.

We compute first and second-order inclusion probabilities for successive sampling and compare them with first and second-order inclusion probabilities of Conditional Poisson and Pareto sampling designs. The successive sampling design was not studied very properly, maybe because it belongs to the same class of order sampling designs with fixed distribution shape as Pareto and Rosén showed that Pareto sampling design is optimal in the class of these sampling designs.

2 Order sampling

Consider a population $U = \{1, 2, ..., N\}$. For each unit i in the population is associated an independent random variable Q_i, called ranking variable, and a probability distribution function F_i, $[0, \infty)$, called order distribution, with density $f_i, i = 1, 2, ..., N$.

Order sampling from population U with sample size n, $n < N$, and order distributions $F_1, F_2, ..., F_N$ is carried as follows. Independent ranking variables $Q_1, Q_2, ..., Q_N$ with distributions $F_1, F_2, ..., F_N$ are realized. The units with the n smallest Q-values constitute the sample.

Let $H(t)$ be a probability distribution function with density $h(t) = H'(t), 0 \leq t < \infty$, and $\theta = (\theta_1, \theta_2, ..., \theta_N)$ are given real positive numbers – intensities. Together $H(t)$ and intensities θ denote the distribution functions $F_i, i = 1, ..., N$.

An order sampling design, $F_i, i = 1, ..., N$, is said to have fixed order distribution shape $H(t)$ with intensities θ, if following two equivalent conditions are met:

1. The ranking variables $Q_1, Q_2, ..., Q_N$ are of type $Q_i = Z_i/\theta_i, i = 1, ..., N$, where $Z_1, Z_2, ..., Z_N$ are independent, identically distributed (iid) random variables with common distribution $H(t)$.

2. The order distributions are $F_i(t) = H(\theta_i t)$, with density $f_i(t) = \theta_i h(\theta_i t)$, $0 \leq t < \infty, i = 1, ..., N$.

Denote $\lambda_1, \lambda_2, ..., \lambda_N$ as target inclusion probabilities for a, maybe approximate, πps sampling design with fixed sample size. Simply $\lambda_1, \lambda_2, ..., \lambda_N$ are given real numbers which satisfy: $0 < \lambda_i < 1, i = 1, ..., N, \sum_{i=1}^{N} \lambda_i = n$. It is shown by Rosén that using the order sampling design with fixed distribution shape, inclusion probabilities π_i can be approximately equal to given target inclusion probabilities $\lambda_i, i = 1, ..., N$.

3 First-order inclusion probabilities

Aires showed that first-order inclusion probabilities for different order sampling designs can be calculated using Lemma 2.

Lemma 2 (Aires, 1999, p. 461). Consider a sequence $Q_1, Q_2, ...$ of independent random variables with distribution functions $F_1, F_2, ...$ Let $Q_{(n)}^N$ be the n-th order statistic among $Q_1, Q_2, ..., Q_N$ with distribution function F_n^N. Then $F_n^N(t)$, $N = 1, 2, ..., n = 1, ..., N$, satisfy recursive equation:

$$F_n^N(t) = F_n^{N-1}(t) + F_N(t) \left(F_{n-1}^{N-1}(t) - F_n^{N-1}(t) \right), \tag{1}$$

where $F_0^N(t) = 1$, for all N and $t > 0$.

In the case of order sampling procedure, the probability of element N belonging to the sample s is:

$$\pi_N = P(N \in s) = P(Q_{(n)}^{N-1} > Q_N) = \int_0^\infty \left(1 - F_n^{N-1}(t)\right) f_N(t) dt. \tag{2}$$

The inclusion probability of any other unit i is derived similarly, from the corresponding formula for the rearranged sequence $Q_1, Q_2, ..., Q_{i-1}, Q_{i+1}, ...,$
Q_N, Q_i instead.

3.1 The successive sampling case

Consider an order sampling design with fixed distribution shape. For successive sampling design the order distribution function can be expressed as $F_i(t) = H(\theta_i t) = 1 - e^{-\theta_i t}, \theta_i > 0$ for $i = 1, ..., N$. The densities then become $f_i(t) = F_i'(t) = (1 - e^{-\theta_i t})' = \theta_i e^{-\theta_i t}$. Parallel to this θ parametrization an alternative set or parameters which are more directly coupled to the inclusion probabilities was used (Aires, 1999): $\lambda_i = F_i(1) = 1 - e^{-\theta_i}, i = 1, ..., N$. This is motivated by the fact that λ_i approximates the inclusion probabilities in case $\sum_U \lambda_i = n$. Let $\tilde{\pi}_i$ denote the inclusion probabilities as functions of λ. Since the intensities, for successive sampling are $\theta_i = H^{-1}(\lambda_i) = -ln(1 - \lambda_i)$, then the probability of element N belonging to the sample s is:

$$\pi_N = -ln(1 - \lambda_N) \int_0^\infty \left(1 - F_n^{N-1}(t)\right) (1 - \lambda_N)^t dt. \tag{3}$$

The exact inclusion probabilities are computed according to Lemma 2, by numerical approximations with a computer program developed with statistical package R. The input of this program is a vector of given target inclusion probabilities $\lambda = (\lambda_1, \lambda_2, ..., \lambda_N)$. At first we compute F_n^N using the recursion in Lemma 2. For numerical integration we use adaptive Simpson′s and Monte-Carlo algorithms. The preprogrammed function for adaptive Simpson′s rule for numerical integration in statistical package R were used.
Example 1. The vector of target inclusion probabilities $\lambda = (0.1, 0.2, 0.3, 0.5, 0.9)$ is given. We compute first-order inclusion probabilities $\tilde{\pi}_i$ using successive sampling design. The population size $N = 5$ and sample size $n = 2$ elements. The control sum is $\sum_{i=1}^{N} \tilde{\pi}_i = 1.999999$, see Table 1.

3.2 The Pareto πps sampling case

Consider an order sampling design and suppose that $F_i(t) = H(\theta_i t) = \theta_i t/(1 + \theta_i t)$ is the standart Pareto distribution function with parameter $\theta_i > 0$ for $i = 1, ..., N$. Then the densities are $f_i(t) = \theta_i/(1 + \theta_i t)^2$. Since $\theta_i = H^{-1}(\lambda_i) = \lambda_i/(1 - \lambda_i)$, then the probability of element N belonging to the sample s is:

$$\pi_N = \lambda_N/(1 - \lambda_N) \int_0^{\infty} \left(1 - F_n^{N-1}(t)\right) \frac{1}{(1 + \lambda_N(t-1))^2} dt. \tag{4}$$

Example 2. We compute first-order inclusion probabilities for Pareto πps sampling design for the same target inclusion probabilities vector given in the example 1, with population size $N = 5$ and sample size $n = 2$ elements. The control sum is $\sum_{i=1}^{N} \tilde{\pi}_i = 1.999999$, see Table 1.

3.3 Conditional Poisson sampling case

Poisson sampling is a method for choosing a sample s of random size $|s|$, from a finite population U consisting of N elements. Each element i in the population has predetermined probability p_i of being included in the sample. A Poisson sample may be realised by using N independent Bernoulli trials to determine whether the element under consideration is to be included in the sample or not. Any experiment that results other that n out of the N elements being selected is rejected. One performs sequentially independent experiments until one of the experiments results in n out of N elements being selected.
First-order inclusion probabilities for conditional Poisson sampling can be calculated using Lemma 1 (Aires, 1999, p. 459).
Lemma 1. Consider a sequence of probabilities $p_1, p_2, ...$ and let $A_n(N)$ be the subset of all samples of size n among $\{1, ..., N\}$ for $n < N$. Then the quantities

$$S_n^N(p_1, ..., p_N) = \sum_{s \in A_n(N)} \prod_{i \in s} p_i \prod_{j \notin s} (1 - p_j)$$

with $N = 0, 1, 2, ...$ and $n = 0, ..., N$, may be calculated recursively by

$$S_n^N(p_1, ..., p_N) = p_N S_{n-1}^{N-1}(p_1, ..., p_{N-1}) + (1 - p_N) S_n^{N-1}(p_1, ..., p_{N-1})$$

for $n = 1, ..., N-1$ using the observations that $S_0^N = (1-p_1)(1-p_2)...(1-p_N)$ and $S_N^N = p_1 p_2 ... p_N$. The inclusion probability $\tilde{\pi}_i$ of any unt $i, i = 1, ..., N$, can be written as:

$$\tilde{\pi}_i = \frac{p_i S_{n-1}^{N-1}(p_1, ..., p_{i-1}, p_{i+1}, ..., p_N)}{S_n^N(p_1, ..., p_N)}. \tag{5}$$

The first-order inclusion probabilities for conditional Poisson sampling design are calculated by a computer program developed with statistical package R. At first S_n^N are calculated using the recursion mentioned above. The input for the program is any vector of unconditional Bernoulli probabilities $p = (p_1, p_2, ..., p_N)$. As a result program returns conditional inclusion probabilities $(\tilde{\pi}_1, \tilde{\pi}_2, ..., \tilde{\pi_N})$.
Example 3. The vector of unconditional Bernoulli probabilities $p = (0.1, 0.2,$
$0.3, 0.5, 0.9)$ is given. We compute first-order conditional inclusion probabilities $\tilde{\pi}_i$, having population size $N = 5$ and sample size $n = 2$ elements. Notice that $\sum_{i=1}^{N} p_i = \sum_{i=1}^{N} \tilde{\pi}_i = 2$, see Table 1.

Table 1: First-order inclusion probabilities for different sampling designs

λ/p	$\tilde{\pi}_i$		
	Successive	Conditional Poisson	Pareto πps
0.1	0.087999779247	0.069470260223	0.094559623047
0.2	0.184249333826	0.154275092936	0.189740430046
0.3	0.290362518450	0.259990706319	0.289828419746
0.5	0.540796307599	0.573187732342	0.517952730787
0.9	0.896591938179	0.943076208178	0.907918436717
sum: 2.0	1.999999877303	2.000000000000	1.9999996403442

4 Second-order inclusion probabilities

Consider an order sampling design with population size N and sample size of n units. Then the bivariate inclusion probability of the units $N-1, N$ is given by:

$$\begin{aligned}\pi_{N-1,N} &= P(N-1 \in s, N \in s) = P(Q_{(n-1)}^{N-2} > max(Q_{N-1}, Q_N)) = \\ &= \int_0^\infty \left(1 - F_{n-1}^{N-2}(t)\right) f_{max(Q_{N-1},Q_N)}(t)dt.\end{aligned} \tag{6}$$

Here

$$\begin{aligned}f_{max(Q_{N-1},Q_N)}(t) &= F'_{max(Q_{N-1},Q_N)}(t) = (F_{N-1}(t)F_N(t))' = \\ &= F'_{N-1}(t)F_N(t) + F_{N-1}(t)F'_N(t).\end{aligned}$$

The inclusion probability of an arbitrary pair of units $i < j$ may be determined by consideration of rearranged sequence $Q_1, Q_2, ..., Q_{i-1}, Q_{i+1}, ..., Q_{j-1}$,
$Q_{j+1}, ..., Q_N, Q_i, Q_j$.

4.1 The successive sampling case

For calculation of second-order inclusion probabilities for successive sampling design we have the same order distribution functions and intensities notations as for the first-order inclusion probabilities. Then second-order inclusion probability for units $i < j$ can be expressed as follows:

$$\pi_{i,j} = \int_0^\infty \left(1 - F_{n-1}^{N-2}(t)\right) f_{max(Q_i,Q_j)}(t)dt, \tag{7}$$

where $f_{max(Q_i,Q_j)}(t) = \theta_i e^{-\theta_i t}(1 - e^{-\theta_j t}) + (1 - e^{-\theta_i t})\theta_j e^{-\theta_j t}$.
Example 4. We compute second-order inclusion probabilities $\tilde{\pi}_{i,j}$ for successive sampling design for the given target inclusion probabilities vector $\lambda = (0.1, 0.2,$
$0.3, 0.5, 0.9)$, where population size $N = 5$ and sample size $n = 2$ elements. The results are shown in Table 2. The control sum is $\sum_{i=1}^{N-1}\sum_{j=1}^{N}\tilde{\pi}_{i,j} = 1$.

4.2 The Pareto πps sampling case

Order distribution functions and intensities for calculation of second-order inclusion probabilities for Pareto sampling are the same as for the first-order ones. In this case second-order inclusion probability for units $i < j$ can be calculated using the equation 7. We give just the expression of the density function:

$$f_{max(Q_i,Q_j)}(t) = \frac{\theta_i}{(1+\theta_i t)^2}\left(1 - \frac{1}{(1+\theta_j t)}\right) + \frac{\theta_j}{(1+\theta_j t)^2}\left(1 - \frac{1}{(1+\theta_i t)}\right).$$

Table 2: Second-order inclusion probabilities for successive sampling design

i	j				
	1	2	3	4	5
1		0.0036335	0.0059265	0.0121894	0.0662504
2			0.0127577	0.0262163	0.1416418
3				0.0426846	0.2289937
4					0.4597060

Example 5. We compute second-order inclusion probabilities for Pareto πps sampling design for the same target inclusion probabilities vector given in the example 4, with population size $N = 5$ and sample size $n = 2$ elements. The control sum is $\sum_{i=1}^{N-1} \sum_{j=1}^{N} \tilde{\pi}_{i,j} = 1$, see Table 3.

Table 3: Second-order inclusion probabilities for Pareto πps sampling design

i	j				
	1	2	3	4	5
1		0.0033026	0.0053607	0.0112222	0.0746742
2			0.0112852	0.0234384	0.1517143
3				0.0374723	0.2357103
4					0.4458199

4.3 Conditional Poisson sampling case

The second-order inclusion probability of units i, j to be included in the sample $s, i \neq j$, can be derived similarly as in the univariate case, using Lemma 1 (Aires, 1999, p. 459) and by consideration of the equations,

$$\tilde{\pi}_{i,j} = \frac{p_i p_j S_{n-2}^{N-2}(p_1, ..., p_{i-1}, p_{i+1}, ..., p_{j-1}, p_{j+1}, ..., p_N)}{S_n^N(p_1, ..., p_N)}. \tag{8}$$

Second-order inclusion probabilities are calculated in the same way as first-order inclusion probabilities. The program was developed with statistical package R. The input for this program is a vector of unconditional Bernoulli probabilities $p = (p_1, p_2, ..., p_N)$. As a result program gives computed second-order inclusion probabilities $\tilde{\pi}_{i,j}$.

Example 6. For given vector $p = (0.1, 0.2, 0.3, 0.5, 0.9)$, second-order inclusion probabilities $\tilde{\pi}_{i,j}$ are calculated, where population size $N = 5$ and sample size $n = 2$ elements. The results are shown in Table 4. Notice that control sum is $\sum_{i<j} \tilde{\pi}_{i,j} = n(n-1)/2 = 1$.

Table 4: Second-order inclusion probabilities for conditional Poisson sampling design

i	j				
	1	2	3	4	5
1		0.0016264	0.0027881	0.0065056	0.0585502
2			0.0062732	0.0146375	0.1317379
3				0.0250929	0.2258364
4					0.5269517

5 Conclusions

Simulation results show that inclusion probabilities for all sampling designs are close, but they do not coincide. We can see that using order sampling design with fixed order distribution shape the exact inclusion probabilities were approximated quite good. The differences can be explained by approximate numerical integration used for calculation of the inclusion probabilities, also by actual differences of those probabilities. The second-order inclusion probabilities differ more than the first-order inclusion probabilities. An approximate integration methods used for calculation of the inclusion probabilities requires long computer execution time .

References

Aires, N. (1999). Algorithms to find exact inclusion probabilities for conditional poisson sampling and pareto πps sampling designs. *Methodology and Computing in Applied Probability* **1:4**, 457 – 469.

Krapavickaitė, D. (2012a). *Implementation of quality improvement actions for the labour force survey.* Statistics Lithuania.

Krapavickaitė, D. (2012b). *Order sampling with fixed distribution shape, calculation of the inclusion probabilities.* Manuscript.

Rosén, B. (1996). On sampling with probability proportional to size. *R&D Report. Research-Methods-Development* **1**, 1 – 25.

Rosén, B. (1998). On inclusion probabilities for order sampling. *R&D Report. Research-Methods-Development* **2**, 1 – 23.

Some remarks on the use of auxiliary information in the Finnish Labour Force Survey

Riku Salonen[1]

[1]Statistics Finland, e-mail: riku.salonen@stat.fi

Abstract

This paper examines the use of auxiliary information both at the sampling stage and at the estimation stage in a complex rotating panel design. Empirical results are based on real data from the Finnish Labour Force Survey (LFS). The motivation for this study comes from the auditing process conducted in 2011 (auditing of the Finnish LFS production system) and recent articles. Many articles evaluate the use of registers as auxiliary information in survey-based statistics and the selection of ”best auxiliary variables” (see e.g. Särndal & Lundström, 2010). The use of several auxiliary information (e.g. administrative registers and published official statistics) both at the sampling stage and at the estimation stage in official statistics production are also discussed by Lehtonen & Djerf (2008) and Fuller (2009). Furthermore Särndal (2007) consider successful application of the use of auxiliary information at the estimation stage in a complex rotating panel design, which part of the information coming from the survey results in previous wave.

Keywords: Complex rotating panel; use of auxiliary information in sampling and estimation

1 Introduction

This paper examines the use of auxiliary information both at the sampling stage and at the estimation stage in a complex rotating panel design. Empirical results are based on real data from the Finnish Labour Force Survey (LFS). The motivation for this study comes from the auditing process conducted in 2011 (auditing of the Finnish LFS production system) and recent articles. Many articles evaluate the use of registers as auxiliary information in survey-based statistics and the selection of ”best auxiliary variables” (see e.g. Särndal & Lundström, 2010). The use of several auxiliary information (e.g. administrative registers and published official statistics) both at the sampling stage and at the estimation stage in official statistics production are also discussed by Lehtonen & Djerf (2008) and Fuller (2009). Furthermore Särndal (2007) consider successful application of the use of auxiliary information at the estimation stage in a complex rotating panel design, which part of the information coming from the survey results in previous wave.

The target population of the Finnish LFS is persons aged 15 to 74, including foreign workers, citizens temporarily abroad, members of the armed forces, non-resident citizens, and unsettled and institutional population. The sample size is approximately 12,000 individuals each month divided into five waves and four or five reference weeks. The monthly sample is allocated so that the weekly sample sizes are equal in each wave. The reference quarters and years are groups of 13 or 52 consecutive weeks. The survey is repeated over time with partially overlapping samples. Each person will be included five times during 15 months. The rotation pattern in the LFS can be described as follows 1-2-1-2-1-5-1-2-1. In the first month, an individual is in the panel in wave one and after a two-month break, he/she will be included in the interview in the second wave, and so on. The lag between the interviews is three months except for one occasion, when it is six months. The design of the LFS ensures the independence of the monthly samples in each three-month period, i.e. a sample for a quarter consists of separate monthly samples. Each sampled person is included once per quarter. This simplifies the estimation of quarterly figures. In the LFS the sample size is 36,000 persons per quarter. There is dependence between successive quarters; the overlap from one quarter to the next is 3/5. There is also a 2/5 overlap between two consecutive years.

2 Sampling stage

The sampling design used in LFS is stratified systematic sampling of elements. The population of individuals is divided into strata. The strata are formed according to NUTS-1 regions (Mainland Finland and The Autonomous Territory of the Åland Islands). In each stratum systematic random selection is applied to the frame sorted according to the domicile code which yields implicit geographic stratification. For estimation purposes, the sampling design is approximated by simple random sampling without replacement (SRSWOR). So far no indication of selection bias due to systematic sampling has been encountered. Systematic sampling can be carried out for example with the SAS procedure SURVEYSELECT.

The sampling frame is based on the database of the total population maintained by Statistics Finland. It is based on the Population Information System of The Population Register Centre and updated regularly. The Finnish system of registers is quite up-to date, especially in register data on individual persons. The updating delay is normally less than one month. The database of the total population is the primary source for information on the population of Finland and it provides the basic information for LFS in both the sampling and estimation stage (gender, age and region). Matching key is PIN.

3 Estimation stage

The current generalised regression (GREG) estimation method was introduced in 1997 and it utilises auxiliary information in the estimation stage (e.g. register data on unemployment). The use of such auxiliary data significantly improved estimates on unemployment by reducing sampling errors and non-response bias (Djerf, 1997). The “register-based job-seeker status” taken from a register maintained by The Ministry of Employment and the Economy’s. This register indicator is classified in different categories e.g. according to the duration of unemployment in the register.

Denote the finite population by $U = \{1,..., k,..., N\}$. A sample $s \subset U$ of size n is drawn by a sampling design $p(s)$ with inclusion probabilities π_k, $k \in U$. Under SRSWOR, the inclusion probabilities are $\pi_k = n/N$. The design weight of unit k is $a_k = 1/\pi_k = N/n$. Denote by y the variable of interest and by y_k its value for unit k.

In the Finnish LFS, post-stratification is used to improve the precision of estimation. The $H = 252$ post-strata are constructed by sex (2 classes), age group (6 groups) and region (21 regions). Let n_h be the number of sampled units in post-stratum h, so $\sum_{h=1}^{H} n_h = n$. At the population level, $\sum_{h=1}^{H} N_h = N$.

There is also missingness due to unit non-response. The weight adjusted for non-response is $d_k = 1/(\pi_k \hat{\theta}_k) = (N_h / n_h) \times (n_h / m_h) = N_h / m_h$ for element k in post-stratum h, where m_h is the number of responding units in post-stratum h and $\hat{\theta}_k = m_h / n_h$ is the estimated response probability for element k in post-stratum h. The weights d_k adjusted for non-response are calibrated using the available auxiliary information. The GREG estimator with linear fixed-effects assisting model is a special case of the calibration estimator (e.g. Särndal, Swensson and Wretman 1992).

As Deville and Särndal (1992 and 1993) show, the GREG estimator of a population total $t_y = \sum_U y_k$ can be given as $\hat{t}_{ygr} = \sum_r w_k^{gr} y_k$ where *r* refers to the respondent group and the calibrated weights are $w_k^{gr} = d_k g_k^{gr}$ with

$$g_k^{gr} = 1 + (\mathbf{t_x} - \hat{\mathbf{t}}_\mathbf{x})' \left(\sum_r \frac{\mathbf{x}_k \mathbf{x}_k ' q_k}{\pi_k \hat{\theta}_k} \right)^{-1} \mathbf{x}_k q_k \,. \tag{1}$$

The known auxiliary totals, called control totals, are $\mathbf{t_x} = \left(t_{x1},...,t_{xj},...\ t_{xJ}\right)'$ and $\hat{\mathbf{t}}_\mathbf{x} = \left(\hat{t}_{x1},...,\hat{t}_{xj},...\ \hat{t}_{xJ}\right)'$ is a vector of estimates of the elements in $\mathbf{t_x}$. The auxiliary information vector is defined as $\mathbf{x}_k = (x_{1k},...,x_{jk},...x_{Jk})'$ and q_k is a known constant (usually set equal to one). The calibration property assures that $\hat{\mathbf{t}}_\mathbf{x} = \sum_r w_k^{gr} \mathbf{x}_k = \sum_U \mathbf{x}_k \ = \mathbf{t_x}$. In the Finnish LFS the auxiliary information vector is defined by four auxiliary variables taken from administrative registers: x_1 = sex (2 classes), x_2 = age (12 groups), x_3 = region (21 regions), x_4 = employment status in job-seeker register (8 classes).

We used a linear distance function in the calibration procedure, available in CLAN, a program developed by Statistics Sweden for calibration and GREG estimation. Variance estimation in CLAN is based on GREG estimation. For variance estimation we need the residuals $e_k = y_k - \mathbf{x}_k' \hat{B}$, where

$$\hat{B} = \left(\sum_r \frac{x_k x_k' q_k}{\pi_k \hat{\theta}_k} \right)^{-1} \sum_r \frac{x_k y_k q_k}{\pi_k \hat{\theta}_k} \,.$$

The variance estimator of $\hat{t}_{ygr}$ under SRSWOR is given by

$$\hat{V}\left(\hat{t}_{ygr}\right) = \sum_{h=1}^{H} \frac{N_h^2}{m_h} \left(1 - \frac{m_h}{N_h}\right) \frac{1}{m_h - 1} \left[\sum_{r_h} \left(g_k^{gr} \times e_k\right)^2 - \frac{\left(\sum_{r_h} g_k^{gr} \times e_k\right)^2}{m_h} \right], \tag{2}$$

where r_h denotes the respondent set in post-stratum *h*.

3 Some remarks on the use of auxiliary information

3.1 Audit findings, questions and conclusions

We present a summary of some audit findings, questions and conclusions concerning the use of auxiliary information.

1) In the Finnish LFS it is not possible to obtain unbiased estimates on unemployment without using proper auxiliary information.

2) The use of job seeker register data as auxiliary information reduced the gap between the two official unemployment figures.

3) Finland belongs to the so-called register countries. Is there more potential auxiliary information available?

4) During the recent decade the non-response rate in Finnish LFS has increased by close to 25 percent. Many articles discuss non-response adjustment. How to select the best and powerful auxiliary information?

5) The use of job seeker register data as auxiliary information improved the precision of unemployment estimates.

6) The regression composite (RC) estimator extends the GREG estimator by using information from the previous wave in a similar manner as the standard GREG estimator uses auxiliary variables. Since 2000, the RC estimator has been successfully used in the Canadian LFS by Gambino *et al*. (2001) . We have compared the RC estimator to the GREG estimator in the Finnish LFS real data by Salonen (2007). The results are well comparable with results reported from other countries (Chen & Liu, 2002).

References

Chen, E.J. & Liu, T.P. (2002). Choices of Alpha Value in Regression Composite Estimation for the Canadian Labour Force Survey: Impacts and Evaluation. *Methodology Branch Working Paper*, HSMD-2002-005E, Statistics Canada.

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

Deville J.-C., Särndal C.E. & Sautory O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013—1020.

Djerf, K. (1997). Effects of Post-Stratification on the Estimates of the Finnish LFS. *Journal of Official Statistics*, 13, 29—39.

Fuller, W.A. (2009). *Sampling Statistics*, Wiley.

Gambino, J., Kennedy, B., & Singh, M.P. (2001). Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation ja Implementation. *Survey Methodology*, 27, 65-74.

Lehtonen, R, & Djerf, K. (2008). Survey sampling reference guidelines: introduction to sample design and estimation techniques, *Eurostat methodologies and working papers*, Office for Official Publications of the European Communities, Luxembourg

Salonen, R. (2007). Regression Composite Estimation with Application to the Finnish Labour Force Survey. *Statistics in Transition*, 8, 503-517.

Särndal, C.E., Swensson, B. & Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

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

Särndal, C-E. & Lundström, S. (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36, 131-144.

Survey Sampling at Vasyl Stefanyk Precarpathian National University

Svitlana Slobodian[1]

[1]Vasyl Stefanyk Precarpathian National University, Ukraine,
e-mail: slobodian_ s@ukr.net

Abstract

The structure of educational direction ”Statistics” is considered. The teaching program for the course on Survey Sampling at the Vasyl Stefanyk Precarpathian National University is presented.

1 Introduction

Statistical science as a whole and survey sampling, in particular, are the subject of detailed interest in the leading universities of Ukraine. Sample surveys play a vital role in a modern society to provide essential information to the government, to politicians, to business and to the citizens. Now the special training course ”Survey Sampling Methods” is given in the educational programme for the speciality ”Statistics” of the Vasyl Stefanyk Precarpathian National University.

2 General information about speciality ”Statistics”

Vasyl Stefanyk Precarpathian National University is the first university of classic type of Precarpathia. It was set up on the anniversary of Ukraine Independence by the Decree of the President. Ivano-Frankivsk Pedagogical Institute and more than 50 years of its functioning laid the foundation of at that time young educational establishment. For a short period of time of its existence university has grown into the leading center of science, culture and education of Precarpathia. Nowadays university contains 6 faculties and 8 institutes. The Faculty of Mathematics and Information Science was set up in 2002 after the reorganization of the Faculty of Physics and Mathematics, whose educational activity can be traced back to 1940. The Faculty of Mathematics and Information Science consists of 6 departments:

- Department of Mathematical and Functional Analysis;
- Department of Differential Equations and Applied Mathematics;
- Department of Algebra and Geometry;
- Department of Information Science;
- Department of Statistics and Higher Mathematics;
- Department of Information Technology.

The Faculty of Mathematics and Information Science contains 4 directions:

- ”Mathematics”;
- ”Applied Mathematics”;

- "Statistics"(since 2010);
- "Information Science".

Speciality "Statistics" of the direction "Mathematics" was opened in 2005 on the base of the Department of Statistics and Higher Mathematics. Direction "Statistics" was allocated in 2010. Now speciality "Statistics" of the direction "Statistics" has two narrow specializations: "Statistics and Actuarial Science" and "Theory Probability and Mathematical Statistics". Educational system of the direction "Statistics" has 3 levels:

- The first educational level is Bachelor of Statistics with four years of training.
- The second educational level is Specialist of Statistics with one year of training (Statistics speciality).
- The third educational level is Master of Statistics with one year of training (Applied and Theoretical Statistics speciality).

The students of the level "Bachelor of Statistics" have 2 weeks of training (statistical) and 6 weeks of working practice, take a state examination and defend a bachelor's thesis. The students of the level "Specialist of Statistics" have 6 weeks of pedagogical and 8 weeks of working practice, take a state examination or defend a diploma thesis. The students of the level "Master of Statistics" have 5 weeks of working practice, defend of master's thesis. They have working practice in statistical institutions and banks.

Certified Specialists of Statistics and Masters of Statistics can work as statisticians, system analysts, risk experts, managers, specialist on IT-technologies, scientific workers and teachers in the fields of statistics and mathematics in different government and commercial institutions.

3 The teaching program of survey sampling

The course of "Survey Sampling Methods" was introduced at the Department of Statistics and Higher Mathematics in 2005 within the speciality "Statistics". Experience of teaching this course in National Taras Shevchenko University of Kyiv was taken into account.

Special course "Survey Sampling Methods" is an additional course of professional choice for the speciality "Statistics" on the first educational level. The timing of this course is 36 hours of lectures and 36 hours of practical lessons.

The program of "Survey Sampling Methods" course includes the following topics:

- Goals and methods of survey
- General scheme of survey
- Simple random sampling with and without replacement
- Sampling with unequal probability
- Systematic sampling
- Stratified random samples
- Simple cluster samples
- Multistage cluster samples
- Linear regression model
- Variance estimation

- Regression estimation
- Errors in surveys, their sources and Methods of their reduction

The lectures on Survey Sampling are based mostly on the books:

- ”Survey Sampling Technique” by Oleksandr Chernyak (2001),
- ”Survey Sampling Methods” by Victoria Parkhomenko (2001).

Also we are planning to use the book by Vasylyk & Yakovenko (2010), the publication of which was supported by the Visby Program grant ”Swedish-Baltic-Ukrainian-Belarusian Collaboration in Survey Statistics”. It will be very useful in teaching the course on ”Survey Sampling Methods” at the Vasyl Stefanyk Precarpathian National University.
During the practical lessons students have the possibility to process data using different software (Statistica, Mathematica, Excel).

References

Chernyak, O. (2001). *Survey sampling technique.* Kyiv (in Ukrainian).

Parkhomenko, V. (2001). *Survey sampling methods.* Kyiv (in Ukrainian).

Vasylyk, O. & Yakovenko, T. (2010). *Lectures notes on the theory and methods of survey sampling.* Kyiv (in Ukrainian).

On the Potential Use of Administrative VAT Data for Estimating Short-term Output Growth in the UK

Markus Gintas Šova[1], Peter John Broad[2] and Craig Brailsford Orchard[3]

[1]Office for National Statistics, UK, e-mail: markus.sova@ons.gov.uk
[2]Office for National Statistics, e-mail: peter.broad@ons.gov.uk
[3]United Kingdom Statistics Authority, e-mail: craig.orchard@statistics.gov.uk

Abstract

The use of administrative VAT data to partially replace survey data for the estimation of short-term output growth offers the prospect of reducing respondent burden. We examine the challenges involved in the UK context and compare various methods of incorporating the VAT data. We conclude that for some divisions VAT data can be used in the estimation of 1-month output growths.

Keywords: Administrative VAT data, short-term output

1 Introduction

There is a growing interest in the potential use of administrative data to partially replace surveys for the estimation of short-term statistics. One such application is the use of Value Added Tax (VAT) data in the estimation of monthly output. The attraction of VAT data is that they are available for all VAT-registered enterprises at no additional cost either to the enterprises or to the National Statistical Institute. Furthermore, protocol 3 of the UK's Code of Practice for Official Statistics (UK Statistics Authority, 2009) states that "Administrative sources should be fully exploited for statistical purposes, subject to adherence to appropriate safeguards." However, the use of VAT data in the compilation of short-term statistics presents certain challenges not encountered when using business surveys. Depending on regulations governing VAT, these may include issues such as timeliness, collection periods and data quality. This paper presents some initial research into the use of VAT data in the UK context. In the following section we discuss the challenges and ways to address them. The methods for using VAT data are described in section 3. Results are presented in section 4, and we conclude with a discussion of our findings.

2 Complicating Factors

In the UK, enterprises are expected to report VAT to Her Majesty's Revenue and Customs (HMRC) according to one of 16 schedules. These are monthly (1 schedule), quarterly (3 schedules, each with a different set of starting months) and annual (12 schedules, each with a different starting month). Approximately 10% of enterprises report monthly, and 0.2% report using an annual schedule (see Orchard, 2010). Parkin (2010) tested several methods of converting the quarterly VAT data into monthly series, but none were found to be superior to simply apportioning the quarterly figures equally into each month covered by the quarter. We have not explored using VAT data reported to an annual schedule because the suitability of annual data to produce useful monthly series for short-term statistics is highly questionable, and because so few enterprises report to an annual schedule.

HMRC receives 100% of VAT returns within 188 days of the end of the reference period. Returns accounting for 40% of turnover are received within 30 days and 94% of turnover within 40 days of the end of the reference period. Parkin (2010) tested various methods of forecasting mature VAT data and recommends the Holt-Winters method. We shall compare this with using mature VAT data directly.

For VAT data to be used they first need to be matched to reporting units (RUs) on the business register. Most RUs are whole enterprises, but some large multi-site enterprises are split into several RUs for statistical purposes. Orchard (2010) describes how VAT data has been apportioned to RUs for such enterprises.

Finally, VAT data can suffer from data quality issues. Lewis (2012) describes how the VAT data has been cleaned.

3 Methods of Incorporating VAT Data

A selection of methods were used to incorporate cleaned VAT data into the UK's Monthly Inquiry into the Distribution and Services Sector (now superseded by the Monthly Business Survey) covering 2005-2008. Their performances were evaluated for NACE Rev.1.1 divisions. Each method keeps the sample *as-is* for larger enterprises and for complex enterprises. The methods are described below.

Method 1 (replacement):
This method assumes that enterprises below a specified employment threshold are not sampled. The cleaned VAT data is used directly for this part of the population. This method was applied using mature VAT data (referring to 6 months and 12 months before the survey period) and forecast VAT data. For the mature data the threshold used was the lower limit of the fourth employment band (usually 100 employees). For the forecast data thresholds of 100 and 250 employees were used.

Method 2 (rescaling from larger enterprises):
This method is similar to method 1, but the cleaned VAT data are first rescaled by a factor equal to the ratio of the survey estimate for larger enterprises to the VAT total (mature or forecast, as appropriate) for larger enterprises. The ratio was calculated for each combination of month and NACE division. The same thresholds were used as for method 1.

Method 3A (rescaling from reduced sample)
This method assumes that the sample size for enterprises below the employment threshold is reduced, but not to zero. The sample reduction was carried out by stratum and tested on sample sizes of 5%, 10%, 25%, 50% and 75% of the existing stratum sample sizes, subject to each stratum retaining a minimum sample size of 5. This allows the cleaned VAT data to be rescaled by a factor comparing the reduced sample turnovers to the VAT data. The factors were calculated in 3 ways: the median of turnover ratios; the ratio of turnover totals; and the trimmed mean of turnover ratios. These factors were calculated for each combination of stratum (for those strata below the employment threshold) and month. This method was only applied to mature VAT data.

Method 3B (alternative rescaling from reduced sample)
Method 3B is the same as method 3A, except that the cleaned mature VAT data is multiplied by two factors. The first factor compares the reduced sample turnovers from the mature data reference period to the mature data (using either the median of turnover ratios, the ratio of turnover totals or the trimmed mean of turnover ratios). The second factor is the ratio of the reduced sample turnover estimate for the survey period to the reduced sample turnover estimate for the mature data reference period.

4 Results

The motivation for this work has been to find methods (or ideally a single method) of using VAT data to at least partially replace a survey of short-term output, specifically 1-month growth. We shall therefore focus on where the methods have performed well (where they have performed badly it would be clearly inappropriate to implement them). We arbitrarily define performing well as the method producing a root mean square difference (RMSD) of less than 5.0 percentage points for 1-month growths.

Table 1: Root mean square differences (in percentage points) of 1-month growths by method and division where 5.0 or less. The 3 rows under the method 3A columns refer to the factor calculation methods (the median of turnover ratios, the ratio of turnover totals, and the trimmed mean of turnover ratios, respectively).

NACE Rev.1.1 Division	Method 1 (12 month old mature data)	Method 1 (6 month old mature data)	Method 1 (forecast data, employment threshold 250)	Method 1 (forecast data, employment threshold 100)	Method 3A (12 & 6 month old mature data, 5% sample retained)		Method 3A (12 & 6 month old mature data, 10% sample retained)		Method 3A (12 & 6 month old mature data, 25% sample retained)		Method 3A (12 & 6 month old mature data, 50% sample retained)		Method 3A (12 & 6 month old mature data, 75% sample retained)	
50	4.9	4.6		3.3	3.1	3.4	2.6	2.7	**2.1**	**2.1**	**1.9**	**1.8**	**1.8**	**1.7**
											3.9	4.3	3.2	3.3
								4.0		2.6	2.8	**2.0**	**2.0**	**1.9**
51	2.4	2.4			**1.8**	**2.0**	**1.8**	**1.9**	**1.7**	**1.8**	**1.6**	**1.7**	**1.6**	**1.7**
													4.5	4.9
						2.6		**2.4**		**2.1**		**1.9**	4.7	**1.8**
55			4.0	3.6										
60	3.8	3.8		4.8	4.9		4.0	4.5	3.3	3.5	2.9	3.0	2.7	2.8
													4.3	4.2
												3.3		2.9
63	3.0	3.8	3.4	3.2	4.7	5.0	4.4	4.7	3.6	3.8	3.2	3.3	3.0	3.1
													3.9	4.3
												4.4		3.6
64	**1.8**	**2.4**	**1.5**	**1.3**	4.6		3.8		2.7	4.3	**2.3**	3.3	**2.1**	2.9
														4.8
71	4.6	4.5	4.5	3.7										
													5.0	4.8
72	4.0	4.0						4.5	3.3	3.7	2.7	3.0	**2.5**	2.7
										5.0		3.4		2.8
73	3.4	3.0	4.7	3.8										
													5.0	
74														
80	**2.4**	**2.4**	**2.5**						4.7		2.9	3.1	**2.4**	**2.4**
85	3.0	**2.1**												
90	2.7	2.6									3.2	3.1	2.7	**2.4**
											4.2	4.5	3.0	3.2
													3.4	4.1
92	4.6	4.7	4.4	3.9										
93											4.6	4.9	3.9	3.9

Based on the definition of performing well, method 1 outperforms method 2 for all divisions using forecast data and 6-month old mature data, and for nearly all divisions for 12-month old mature data. Method 3B performs poorly nearly everywhere. Table 1 shows the RMSDs for the other methods where they are 5.0 percentage points or less, with RMSDs of 2.5 percentage points or less are shaded. It is immediately clear that no single method is best for all divisions, although method 1 appears to do well for most. For method 3A, the median of turnover ratios generally provides the best factor for rescaling the VAT data. No method has been found that works well for division 74.

5 Discussion

This project has demonstrated that VAT data can be used to partially replace surveys for measuring short-term (1-month) output growth for some divisions. However, the most appropriate method to use varies by division and for at least one division no suitable method has been identified. Broad (2012) extends this analysis to cover 12-month output growths and output levels, which any recommendations for implementation will need to take into account. Our arbitrary definition of good performance, a root mean squared difference of no more than 5.0 percentage points, may at first sight seem large, but it should be regarded in the context of the standard errors of the existing survey at division level.

Once recommendations are made as to which methods (if any) to use for which divisions, the impact at the top level of aggregation will need to be evaluated. Further research should be based on NACE Rev.2 and be extended to include the production industries. Finally, it is thought that there is some scope for a more refined apportionment of quarterly VAT returns into months.

Acknowledgements

The authors are grateful to ONS colleagues Ann Langford and Mila Teneva for their important contributions to this work. Part of this work was carried out under work package 4 of the ESSNet project “The Use of Administrative and Accounts Data for Business Statistics.”

References

Broad, P.J. (2012). *Use of administrative VAT data in the Monthly Business Survey.* ONS internal paper, available from the author on request.

Lewis, D.J. (2012). *Cleaning VAT Turnover data for use with MBS mixed-source estimates.* ONS internal paper, available from the author on request.

Orchard, C.B. (2010). *File Preparation, Data Matching, and Distribution Analysis of VAT Turnover.* ONS internal paper, available from the author on request.

Parkin, N. (2010). *Interpolation and Extrapolation from Value Added Tax Returns.* ONS internal paper, available from the author on request.

UK Statistics Authority (2009). *Code of Practice for Official Statistics,* edition 1.0. UK Statistics Authority, available at:
http://www.statisticsauthority.gov.uk/assessment/code-of-practice/code-of-practice-for-official-statistics.pdf

Finite mixtures analysis by biased samples

Olena Sugakova[1] and Rostyslav Maiboroda[2]

[1]Kyiv National University, Ukraine, e-mail: sugak@univ.kiev.ua
[2]Kyiv National University, Ukraine, e-mail: mre@univ.kiev.ua

Abstract

We propose new estimates for means and CDFs of finite mixture components when the mixing proportions are not constants and some samping bias is present.

Keywords: Horwitz-Thompson sampling bias correction, weighted empirical cumulative distribution function, mixture with varying concentrations, finite mixture model

1 Introduction

In this presentation we discuss means and CDFs estimation of finite mixture components in the case when the observed sample is subjected to some sampling bias. The proposed estimation technique borrows from the Horwitz-Thompson (HT, see Lohr, 2010, p. 241) bias correction and the methodology of mixtures with varying concentrations (MVC) analysis, see Maiboroda & Sugakova (2012). In what follows we present a motivating example (Section 2), describe the HT estimates by a homogeneous sample (Section 3) and the MVC estimates by unbiased samples (Section 4). Then the estimates by both mixed and biased data are introduced in Section 5. Results of simulations are presented in Section 6.

2 Motivating example

Imagine that the distribution of some characteristic ξ (e.g. body length) of crabs living at a sea is studied. The investigated population of crabs is divided into two sub-populations (components). The crabs belonging to the first component are more salt-loving then the ones belonging to the second component. Both components are present in all sites in the sea where the crabs are caught, but their proportion in the local population depends on the mean salinity of the water at this site. Assume that the function describing the dependence of this proportion from the salinity is known.

We are interested in the differences in distribution of ξ for crabs belonging to different components. But the true component to which the crab belongs is not observed in the study, since it needs some expensive and time consuming tests. Therefore our inference should be based on the proportions of components at the sites where the crabs were caught. These proportions can be considered as probabilities that a crab chosen at random from the local population belongs to a given component (mixing probabilities). They can be estimated by the mean salinity data.

To catch the crabs some traps are used and it is known that the probabilities to be caught are different for crabs with different body length ξ. This causes a sampling bias in the observed distribution of ξ. Our aim is to correct this bias and to extract the CDF of the component of interest from the mixture.

3 Horwitz-Thompson approach to bias correction

Assume that there is a homogeneous infinite (very large) population of subjects O with the observed feature $\xi(O) \in \mathbb{R}$. The CDF of $\xi(O)$ in the entire population is $F(x)$. A subject O can be sampled from the population with the probability depending on $\xi(O)$ but independently of all other subjects. (See Shao, 2003, p. 328). Let us denote this (inclusion) probability by

$$cq(t) = \mathsf{P}\{O \text{ was included to the sample } |\xi(O) = t\}, \tag{1}$$

where $q(t)$ is a known function, c is an unknown constant. The values of $\xi(O)$ for the sampled subjects constitute the sample $Y = (\eta_1, \eta_2, \dots, \eta_n)$. It is obvious that η_i are i.i.d. and the their CDF $\tilde{F}(x)$ is the conditional probability of the event $\{\xi(O) < x\}$ given that O was sampled. Then

$$\tilde{F}(x) = \mathsf{P}\{\xi(O) < x \mid O \text{ was sampled}\} = \frac{\int_{-\infty}^{x} q(t)F(dt)}{\int_{-\infty}^{+\infty} q(t)F(dt)}. \tag{2}$$

In this case the population mean $\bar{\xi} = \mathsf{E}\,\xi = \int xF(dx)$ does not equal to the expectation of the observed values $\mathsf{E}\,\eta_j$ due to the sampling bias. But $\bar{\xi}$ can be estimated by the weighted sample mean with weights reciprocal to the inclusion probabilities:

$$\hat{\xi} = \frac{1}{\sum_{j=1}^{n} \frac{1}{q(\eta_j)}} \sum_{j=1}^{n} \frac{1}{q(\eta_j)} \eta_j.$$

It is the usual HT-estimate which is known to be consistent if $\bar{\xi}$ exists and $q(t) > \mathit{const} > 0$ for all t. The corresponding estimate for CDF F is

$$\hat{F}^{HT}(x) = \frac{1}{\sum_{j=1}^{n} \frac{1}{q(\eta_j)}} \sum_{j=1}^{n} \frac{1}{q(\eta_j)} \mathbb{1}\{\eta_j < x\}.$$

4 Mixtures with varying concentrations

In the MVC model we assume that the subjects can belong to one of M sub-populations (components) $\mathcal{P}_1$, $\mathcal{P}_2$,…,$\mathcal{P}_M$. The probability to observe a subject from a given component depends on the conditions of the observations and is, generally speaking, different for different observations. Let us denote by p_j^i the probability to observe a subject from $\mathcal{P}_i$ in the j-th observation. The CDF of the observed feature $\xi(O)$ of a subject O is different for different components: $H_m(x) = \mathsf{P}\{\xi(O) < x \mid O \in \mathcal{P}_m\}$. So, the observed sample X consists of independent but not identically distributed observations $X = (\xi_1, \dots, \xi_n)$ with CDFs

$$F_j(x) = \mathsf{P}\{\xi_j < x\} = \sum_{m=1}^{M} p_j^m H_m(x). \tag{3}$$

Note that the component to which an observed subject belongs is unknown. One needs to infer on H_m only by the sample X and the set of mixing probabilities (concentrations) p_j^m, $j = 1, \dots, n$, $m = 1, \dots, M$ which are known (estimated). We do not assume any sampling bias now.

The CDF H_m may be estimated by a weighted empirical CDF

$$\hat{H}_m(x) = \frac{1}{n} \sum_{j=1}^{n} a_j^m \mathbb{1}\{\xi_j < x\}.$$

Here a_j^m are some weights which may depend on mixing probabilities p_j^i, but not on the observations ξ_j.

To obtain unbiased estimates one needs the following conditions to be satisfied:

$$\frac{1}{n} \sum_{j=1}^{n} a_j^m p_j^i = \mathbb{1}\{i = m\} \text{ for all } i = 1, \dots, M.$$

One of appropriate choices is the minimax weighting with

$$a_j^m = \sum_{i=1}^{M} \bar{\gamma}_{im} p_j^i,$$

where $\bar{\Gamma} = (\bar{\gamma}_{im})_{i,m=1}^M$ is the matrix inverse to $\Gamma = (\frac{1}{n}\sum_{j=1}^n p_j^i p_j^k)_{i,k=1}^M$.

To estimate the m-th component mean $\bar{\xi}_m = \int x H_m(dx)$ one may use

$$\hat{\xi}_m = \frac{1}{n}\sum_{j=1}^{n} a_j^m \xi_j.$$

5 Mixture model with sampling bias

Let us assume now that the MVC model is in force together with the sampling bias. It means that the considered subjects O belong to M different components and the CDF of their feature of interest $\xi(O)$ is H_m for the subjects belonging to the m-th component. The proportion of the m-th component subjects in the local population from which the j-th subject was obtained is p_j^m. These probabilities are known. The CDFs H_m are unknown. The sampling is biased in the sense that the probability to sample the subject O from a local population depends on $\xi(O)$. This probability is defined by (1) with known q and unknown c.

The problem is to estimate the components' CDFs H_m and means $\bar{\xi}_m = \int x H_m(dx)$.

Analogously to (2) we obtain that the observed sample $Y = (\eta_1, \dots, \eta_n)$ consists of independent observations with CDFs

$$\tilde{F}_j(x) = \mathsf{P}\{\eta_j < x\} = \frac{\int_{-\infty}^{x} q(t) F_j(dx)}{\int_{-\infty}^{+\infty} q(t) F_j(dx)}, \tag{4}$$

where F_j is defined by (3).

From (4) we obtain

$$\tilde{F}_j(x) = \sum_{m=1}^{M} \frac{p_j^m \tilde{Q}_m}{\sum_{i=1}^M p_j^i \tilde{Q}_i} \tilde{H}_m(x),$$

where $\tilde{Q}_m = \int_{-\infty}^{\infty} q(t) H_m(dx)$, $\tilde{H}_m(x) = \int_{-\infty}^{x} q(t) H_m(dx)/\tilde{Q}_m$. So, the sampling bias causes changes not only in the distributions of components, but also in the mixing probabilities. To take in account the bias in the mixing probabilities, we need to estimate $\tilde{Q}_m$. This can be done by observing that

$$\mathsf{E}\frac{1}{q(\eta_j)} = \frac{1}{\sum_{i=1}^M p_j^i \tilde{Q}_m}.$$

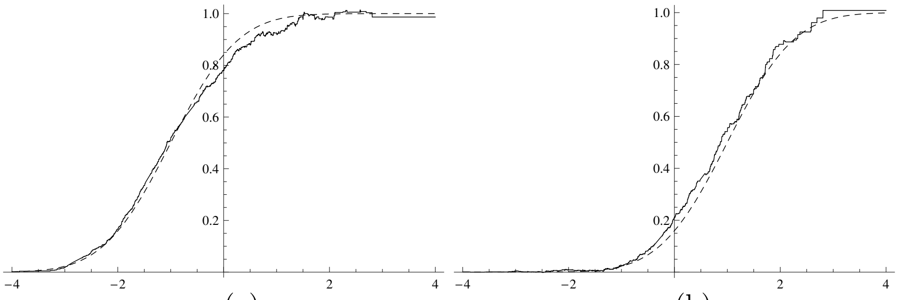
Then the least squares technique suggests the estimate $\hat{\mathbf{Q}} = (\hat{Q}_1, \dots, \hat{Q}_M)$ for $\tilde{\mathbf{Q}} = (\tilde{Q}_1, \dots, \tilde{Q}_M)$ which is the minimizer of the LS functional

$$J(\mathbf{Q}) = \sum_{j=1}^{n} \left(\frac{1}{\sum_{i=1}^M p_j^i Q_i} - \frac{1}{q(\eta_j)} \right)^2$$

over all $\mathbf{Q} = (Q_1, \dots, Q_M)$ with $Q_i > 0$.

With these estimates at hands we define the weights $\tilde{a}_j^m$ for the m-th component estimation as

$$\tilde{a}_j^m = \frac{1}{q(\eta_j)} \sum_{k=1}^{M} \tilde{\gamma}_{km} \frac{p_j^k}{\sum_{i=1}^M p_j^i \hat{Q}_i},$$



(a) (b)

Figure 1: Estimates of CDF for the first (a) and second (b) component. The true CDFs are depicted by dashed lines.

where $\hat{\Gamma}_Q = (\hat{\gamma}_{km})_{k,m=1}^M$ is the matrix inverse to

$$\tilde{\Gamma}_Q = \left(\frac{1}{n} \sum_{j=1}^n \frac{p_j^k p_j^m}{\left(\sum_{i=1}^M p_j^i \hat{Q}_i \right)^2} \right)_{k,m=1}^M .$$

The resulting bias-correcting estimate for the H_m is

$$\hat{H}_m^{BC}(x) = \frac{1}{n} \sum_{j=1}^n \tilde{a}_j^m \mathbb{1}\{\eta_j < x\}.$$

The estimate for $\bar{\xi}_m$ is

$$\hat{\xi}_m^{BC} = \frac{1}{n} \sum_{j=1}^n \tilde{a}_j^m \eta_j.$$

6 Results of simulation

We performed a small simulation study to assess performance of the proposed estimates. In our experiment we considered a two component mixture $M = 2$. The distribution of the first component was $N(-1, 1)$, the distribution of the second one was $N(1, 1)$. The concentrations of the first component p_j^1 were simulated as random variables, uniformly distributed on $[0, 1]$, $p_j^2 = 1 - p_j^1$. Figure 1 presents the graphs of the estimates $\hat{H}_m^{BC}(x)$ for the components CDFs by a sample with $n = 1000$ observations.

The biases and variances of the estimates $\hat{\xi}_m^{BC}$ for different sample sizes n are presented in Table 1.

Table 1: Performance of the estimates for means

n	ξ_1^{BC}		ξ_2^{BC}	
	bias	Var	bias	Var
50	-0.0807	0.8011	-0.0851	0.3360
100	-0.0581	0.2257	-0.0589	0.1575
250	-0.0176	0.0849	-0.0271	0.0815
500	-0.045	0.0464	0.00210	0.0330
750	-0.0285	0.0421	-0.0162	0.0187
1000	-0.0052	0.0211	-0.0034	0.0118

7 Concluding remarks

We have constructed the estimates for means and CDFs of mixture components in the case when the sampling procedure is biased. The simulations indicate satisfactory behavior of the estimates in the case of Gaussian mixture. More efforts are needed to analyze the asymptotic behavior of these estimates and their performance on non-Gaussian data.

References

Lohr, S. (2010) *Sampling: Design and Analysis.* Brooks/Cole.

Maiboroda, R. & Sugakova, O.(2012) *Statistics of mixtures with varying concentrations with application to DNA microarray data analysis.* Nonparametric statistics, **24**, iss.1, 201-215.

Shao, J. (2003) *Mathematical statistics.* Shpringer.

Combination of sample surveys or projections of political opinions

Daniel Thorburn[1] and Can Tongur[2]

[1]Stockholm University, e-mail: Daniel.Thorburn@stat.su.se
[2]Stockholm University and Statistics Sweden, e-mail Can.Tongur@scb.se

Abstract

In Sweden sample surveys of the party preferences are made almost every month by different institutes. The sample sizes are usually between 1000 and 2000 which means that the standard deviations are between 1 and 1.5 %. We study how these estimates can be combined to get better estimates taking the trends and voter mobility into account. Our model is a combination of a dynamic model based on Wienerprocesses and sampling theory with design effects. Since the party preferences are modelled as random processes it will be possible also to talk about the probability for events like a party (block) has more than 50 % of the preferences. Assuming that the same model and the same parameters will hold also in the future we can also give intervals for the future election results. short abstract is nice at the beginning of the text to describe the contents and main results of the paper.

Keywords: Dynamic models, Metaanalysis, Party preferences, Poll of polls, Sample surveys, Wiener processes

1 Introduction

In Sweden many different institutes make opinion polls almost every month . Four important private actors are SIFO, Temo, SKOP and Novus. The sample sizes are usually between 1000 and 2000. Statistics Sweden makes a poll twice a year with a sample size of about 7000. We study how these estimates can be combined to get better estimates taking the trends and the voter mobility into account. Our approach is a combination of a dynamic model based on Wienerprocesses (West & Harrison, 1997) and sampling theory with design effects. Assuming that the same parameters will hold also in the future it will be possible to give prediction intervals for the upcoming election results.

A recently presented method to weight together previously presented polls is "Poll of polls". A simple description is given by Salmond.(2012). A more detailed description is given by Silver (2008). His basic idea is to weight the previous polls in order to get the best estimate of the present opinion. Silver (2008) also discussed estimation of trends. He used a simple Gibbs sampling technique based on regression analysis to forecast the outcome of the next election. We suggest the use of a random walk process as the underlying model to avoid the rigidity of linear models.

2 Simple Basic Model

In our basic model the proportion of people supporting a party/party block preferences behaves like a random walk over time. This is a mathematical model and does not explicitly take all available extra information into account. For example, will the time when a party changes its leader be modeled as taking place at an unknown time in the future random which cannot be exactly predicted. In the same way all other important events like financial crises, sexual scandals, political debates or special campaigns are viewed as random events which cannot be predicted neither the time nor its effects. The model uses only the observed data and does not use subjective but generally held beliefs as the observation that the opinion usually swings against the sitting government in the middle of an parliamentary term to go back when the election gets closer.

In this random walk model it will be equally likely that a party decreases or increases from the present level. What happens will depend on future unknown events. A specialist in political science can very likely improve on the model by using his expert knowledge. Our aim, however, is making estimates and predictions using only observed polls. The proportion of a party/party block at time t is denoted P_t and its development is modelled by a random process given by the stochastic differential equation

$$dP_t = \gamma \, dW_2(t),$$

where $W_2(t)$ is white noise and contains all the small and large things that affect the voters' opinions. The solution to this equation says that if a party at a certain day t has the proportion P_t, than the distribution s days ahead will be normally distributed around P_t and with the variance γs (and thus standard error $\sqrt{(\gamma s)}$). The proportion is usually unknown but measured with a random error by opinion polls and exactly by general elections.

Suppose that the model says that P_t is normal with mean EP_t and variance CPP_t at a certain time. Suppose further that at time t the result X_t of an opinion poll becomes available measuring P_t with the variance V_t. Combining the prior belief with the observed proportion the best guess of the true proportion P_t is that it is normally distributed with a mean that is weighted between the prior mean and the observed poll. The weights should be inversely proportional to the variances

$$\frac{V_t * EP_t + CPP_t * X_t}{V_t + CPP_t}. \qquad (1)$$

The new variance is given by the smaller value

$$\frac{V_t * CPP_t}{V_t + CPP_t}. \qquad (2)$$

This model is very similar to the models used by financial experts modelling prices at the Stock Exchange. There is a lot of information around but most of the effects are already capitalised in the prices and it is difficult to predict the future development of stock prices. It may be possible but the fact is that only very few persons succeed in making a fortune on the stock market shows that most of the information is already reflected in the prices. It may also be viewed as a Dynamic Linear Model using Kalman filters (West & Harrison, 1997).

3 A Model with Trend, Forecasting s Periods Ahead

3.1 Background

Here we will introduce a trend into the simple model above. Thus the best predictor of the future given only the previous polls was earlier the same as the best estimate of the present level. One might argue that there can be a trend. If a party has been steadily increasing its share of the voters for the last period one could expect that it would continue to do so. To study this, a model with trend is introduced. But the trend cannot be going on forever so the chosen model says that the trend is expected to decrease gradually and be replaced by other trends. The model contains three parameters, γ, which measures the size of the short term fluctuations, α, which measure how fast the trend disappears (e.g. = 0.05 means half the trend will have disappeared after roughly 1/0.05=20 days) and β, which measure how much randomness is explained by the (changing) trend.

3.2 Formulas

Let P_t, as before, be the true level of the sympathy for a party (party block) and let T_t be the trend in P_t.

Assume that the trend behaves like Ornstein-Uhlenberck process, i.e. it follows the stohastic differential equation

$$dT_{t}\_ = -\alpha T_t dt + \beta dW_1(t),$$

where dW_1 is white noise and α and β are positive constants. Solving and expressing T_{t+s} in terms of T_t gives

$$T_{t+s} = \exp(-\alpha s)T_t + \beta\int_t^{y+s} \exp(-\alpha(t+s-u))dw_1(U). \tag{3}$$

The party's share of the votes has a drift proportional to T_t and a constant noise term at time t, i.e. it follows the stochastic differential equation

$$dP_t = T_t + \gamma\, dW_2(t)$$

where dW_2 is another independent white noise and γ a positive constant. Solving and expressing the future party sympathy in terms of the situation at time t gives

$$P_{t+s} = P_t + \int_t^{t+s} T_u du + \gamma\int_t^{t+s} dW_2(u) \tag{4}$$

If the expression (3) for T_t is inserted we get that P_{t+s} equals

$$P_t + \int_t^{t+s} (\exp(-\alpha(u-t)T_t) + \beta\int_t^u \exp(-\alpha(u-v))dW_1(v))du + \int_t^{t+s} dW_2(u) \tag{4}$$

The first part of the middle term is easily computed. For the second part we change the order of integration and after that compute the (new) inner integral. The result is that P_{t+s} equals

$$P_t + ((1+\exp(-\alpha s)/\alpha)T_t + \beta\int_t^{t+s} (1-\exp(\alpha(t+s-v)))/\alpha + dW_1(v) + \gamma\int_t^{t+s} dW_2(u) \tag{5}$$

The next problem is to update the expected values and variances when a new opinion poll is observed. Before the observation at time t the expected levels are denoted EP_t and ET_t and the variances CPP_t, CTT$_t$_and CPT$_t$. After observing X$_t$ from a poll with precision Var(X$_t$ -P$_t$) = V$_t$, the prior is updated by combining the observation and the prior getting a posterior. The expected levels are updated exactly as in Formulas (1) and (2).

$$E(P_t \mid X_t) = \frac{V_t * EP_t + CPP_t * X_t}{V_t + CPP_t}.$$

with the variance

$$Var(P_t \mid X_t) = \frac{V_t * CPP_t}{V_t + CPP_t}.$$

Updating the distribution of the trend is also fairly simple.

4 Other Features

4.1 Design Effects

Even though the four Swedish institutes use slightly different techniques, the methods are fairly similar. There are also studies based on web surveys in Sweden but we will not use any of them. Their results are not completely comparable to those based on probability sampling. The party preference study of Statistics Sweden is based on a simple random sample from the Swedish population register and is performed by telephone interviewing with about 30% non-response. About half of the non response consists of refusals and the other half are persons who are not found (e.g. not at home or no telephone number found). The studies by the private institutes are based on some version of RDD and telephone interviewing. All institutes use some sort of weighting to decrease the variance. Statistics Sweden calibrates with some known register variables. Most institutes ask about the voting at last election and weights to some extent with e.g age, sex and the outcome at that election. This means that there probably is a design effect and that the variance is smaller than what the binomial distribution formula says (1/(nP$_t$ (1-P$_t$))).

We will assume that the design effects are the same for all institute gives

$$\begin{aligned} Var(X_t - P_t \mid P_t) &= V_t \\ &= \delta P_t (1 - P_t) / n \end{aligned} \qquad (4)$$

(The data does not allow us to efficiently estimate the different design effects without introducing an informative prior). The institutes have also different ways of formulating the interview questions. Thus we allow them to have different biases.

4.2 Parameter Estimation

The model described above contains many parameters, α, β, γ, δ and the institute effects. The model also contains starting values, i.e. the means EP$_0$ and ET$_0$ and the variances CPP$_0$,CTT$_0$and CPT$_0$. Since the data start at an election with the outcome X$_0$, the choice of starting values are natural. As we are going to study a longer period our solutions will be quite robust against miss-specified starting values.

In order to estimate the parameters owe maximise the likelihood function. Since each observation of X$_t$ - EP$_t$ given the history (i.e. all observations before time t) is approximately normal, the following likelihood can be used

$$L(\alpha,\beta,\gamma,\delta) = \prod_t (1/((2\pi(V_t+CPP_t))^{1/2})) \exp(-(((X_t-EP_t)^2)/(2(V_t+CPP_t)))),$$

where the product is over all opinion polls. Taking logarithms this becomes.

$$-2l(\alpha,\beta,\gamma,\delta) = \sum_t ((X_t - EP_t)^2)/(V_t + CPP_t) - \ln(V_t + CPP_t)$$

Note that the last term is important since it depends on the parameters. This expression can also be used for testing hypothesis on the parameters.

4.3 Variance Stabilising Transformation

Whenever the proportion of voters for a certain party or party block is relatively large, it may be reasonable to assume that its development can be described by the processes described above. However, when the proportions are close to zero (or one) the process may reach zero and the forecasted proportions may become negative. To cope with this situation, which is likely to occur in an election system having several smaller parties, we use a logistic transformation

$$Q_t = \ln(P_t) - \ln(1-P_t)$$

and assume that Q_t will follow the above processes. The logistic transformation has also the advantage that the process will never become larger than 1. Additionally, this model is symmetric in the sense that it can be used both for the party size and also the complimentary event of not voting for this party (with opposite sign).

If X_t is the result of a poll at time t with mean P_t and variance V_t, then $Y_t = \ln(X_t) - \ln(1-X_t)$ is a an estimator with approximate mean Q_t and approximate variance $V_t/(P_t(1-P_t))^2$. If the variance is given by (6) this simplifies to

$$Var(Y_t - Q_t) = \delta/(nP_t(1-P_t))$$

Thus we can use the same formulas as above for this process but use this variance expression in the innovation formulas.

5 Data

5.1 Parties in Sweden

The parties in the Swedish Parliament can be divided int0 two blocks. The Bourgeois block called the Alliance forms the government since 2006 and comprises the conservatives, "Moderaterna" (M), the christan democrats (Kd), the liberal party, "Folkpartiet" (Fp), and the centre party (C). The opposition consists of the three red or green parties; They are the green party, "Miljöpartiet" (Mp), Labour, "Socialdemokraterna" (S), and the left party, "Vänstern" (V). During half of the last parliamentary trem these three term formed a coalition with the object of winning the election in 2010 and forming the government together. However they did not gain power and now they act as three separate parties. However, it is still common among political commentators, to compare their total size to that of the alliance. These seven parties were represented in in the parliament in 2006. In 2010 a new populist party, Sverigedemokraterna (Sd), took seats in the parliament but does not belong to any of the two blocks. There exist also some other small parties in Sweden, which are not represented in the parliament.

5.2 The Data Set

For the analysis in this paper we have used all Swedish party preference studies from October 2010 until beginning of March 2012, altogether some 258 observations from five institutes. We have focused on the four established private institutes Novus, SIFO, Skop and Temo/Synovate/Ipsos but also Statistics Sweden's opinion polls as well, together with the election result of 2010. The data are from the public home page of the Novus group (2012). In this paper we focus on the proportion of Bourgeois voters of all parties in the parliament at that time.

The fit of the model is shown in the first graph. Here the results of all opinion studies are shown together with 95 % prediction intervals given all the previous studies. (When two studies have the same reference date only the interval for the last study is shown). The raggedness of the boundaries depends partly on the fact that the studies had different sizes and that the prediction intervals become shorter for large studies. Another reason is that each time that new information is added the predicted level changes with a small jump.

6 Estimation of the Party Preferences

The previous discussion concerned the prediction of the opinion polls given all previous data. The intervals were ragged due to the different sizes of the polls. The second graph shows the same period but now we estimate the true level for the period and use both old and future measurements.

Even though our goal with this project was meta-analysis, the model can be used for mechanical projections. The last graph shows forecasts for the election in 2010 at different times after the election in 2006, given the polls that have been published at that data.

References

Eklund J. & Järnbert, M. (2011), The Party Preference Study (PSU) – Description of the Statistics, ME0201, (in Swedish), SCB, Stockholm.

Novus Group, (2012), All Swedish Opinion Polls, to be found at http://www.novusgroup.se/vaeljaropinionen/samtliga-svenska-vaeljarbarometrar, (apr 2012)

Salmond, R., (2012), Pundit Poll of Polls: How we do it, http://pundit.co.nz/content/poll-of-polls, (feb 2012), *Pundit*

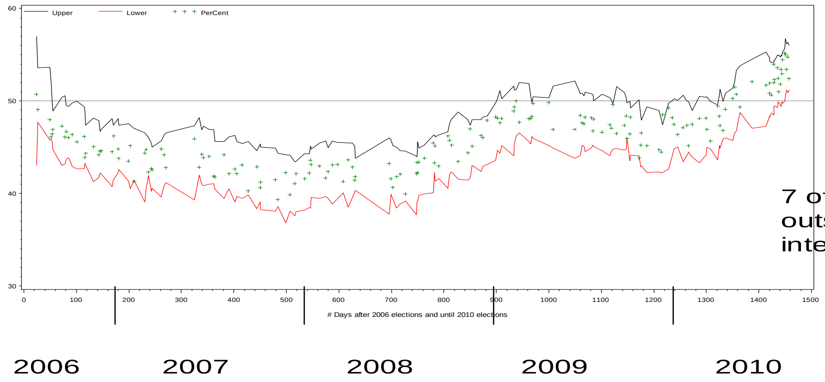
Silver, N., (2008), Politics done right, http://www.fivethirtyeight.com/2008/03/frequently-asked-questions-last-revised.html , (feb 2012) *Fivethirtyeight*

West, M. & Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, 2nd edit, Springer

Figure 1 The alliance, prediction of polls

(based on data before the poll, interelection period)

P(institute) and prediction intervals around the prior



7 of 175 outside the interval

Estimates for the Alliance during last interelection period

All data, 95 % probability intervals, excl. minor parties and Sd

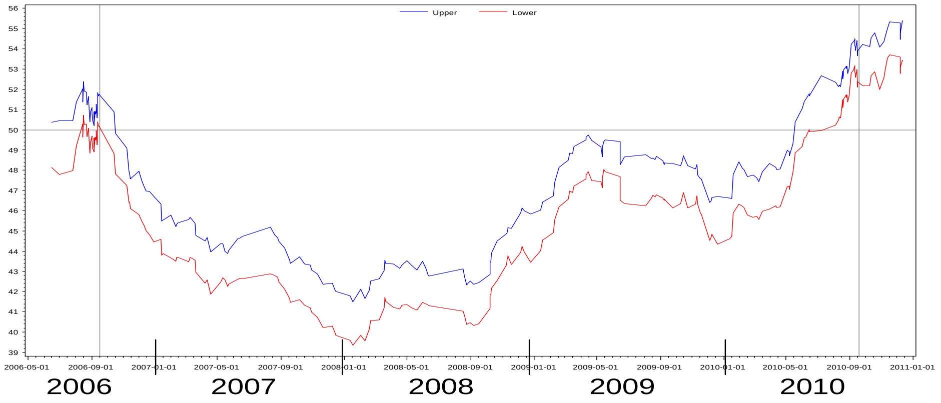
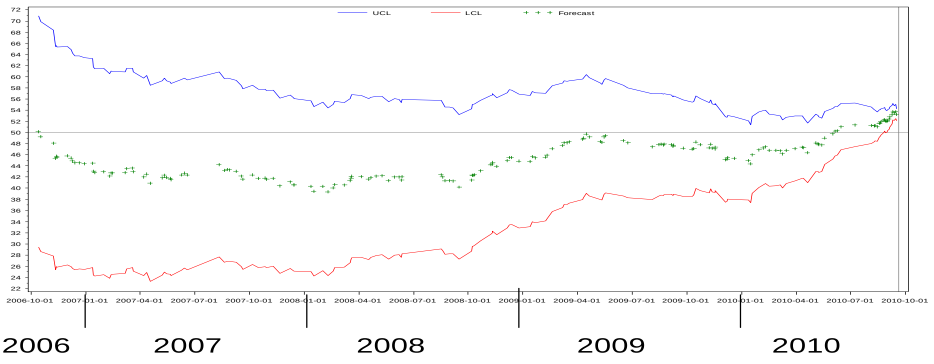


Figure 3 Forecasts for the 2010 election, based on previous polls



The number of Latvian residents estimation via logistic regression

Jeļena Vaļkovska[1]

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

Abstract

There were 2 070 371 residents of Latvia on 1 March 2011 according to the results of Population and Housing Census 2011. Further data processing and routine statistics production will be based on the Population and Housing Census 2011 results. Therefore, it is very important to know the number of the residents in some fixed moment of time. It was decided to investigate the logistic regression as the potential instrument. The main aim of this research is to get such estimates of beta coefficients, that can be used to estimate the number of Latvian residents and number of emigrants now and in the future.

Keywords: census, logistic regression, residents, estimators, balanced sampling

1 Introduction

The Population and Housing Census was organised in March 2011 in compliance with the Law on Population and Housing Census and European Parliament and Council Regulation (EC) No 763/2008 of July 9, 2008. The main aim of it was to obtain detailed enough view on structure and characteristics of the population.

Resident population comprised all persons who:

1. have lived in Latvia for at least 12 months before the Census moment (01.03.2011);
2. have arrived in Latvia within 12 months before the Census moment with an intention to spend at least one year in the country.

Persons not counted in the Census:

1. registered at the Population Register, but residing outside of the territory of Latvia for more than 12 months;
2. have entered the country less than 12 months before the Census moment, are residing in the country but are not planning to stay in Latvia for more than 12 months;
3. individuals born after 01.03.2011;
4. individuals who have died before 01.03.2011;
5. foreign armed forces, sea force and consular personnel and their family members residing the country;
6. tourists. (CSB home page: Population Census)

2 Logistic regression

The aim of an analysis using logistic regression is the same as that of any model-building technique used in statistic: to find the best fitting and most parsimonious model to describe the relationship between dependent variable and a set of independent variables. What distinguishes logistic regression from the linear regression model is that the outcome variable in logistic regression is dichotomous.

Let the vector of t independent variables is $x^{'} = (x_1, x_2, ..., x_t)$. The quantity (1) is used in multiply regression to represent the conditional mean Y given x and quantity (2) is the logits transformation.

$$\pi(x) = E(Y|x) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_tx_t}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_tx_t}} \tag{1}$$

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = e^{\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_tx_t} \tag{2}$$

The logit $g(x)$ has many of the desirable properties of a linear regression model: it is linear in its parameters, may be continuous, and may range from $-\infty$ to $+\infty$, depending on the range of x.

The model fitting means that it will be obtained the estimates of vector $\beta^{'} = (\beta_1, \beta_2, ..., \beta_t)$. The method of estimation used is the maximum likelihood. The main principle of this method is that is used as the estimate of β the value which maximizes the expression:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left[1-\pi(x_i)^{(1-y_i)}\right]. \tag{3}$$

The comparison of observed to predicted values using the likelihood function is based on the expression:

$$D = -2ln\left[\frac{likelihood\ of\ the\ fitted\ model}{likelihood\ of\ the\ saturated\ model}\right]. \tag{4}$$

Probabilities are always less than one, so the value of *log likelihood* (LL) is always negative, that is why the expression is multiplied by (-2). Such test is called likelihood ratio test:

$$D = -2ln\sum_{i=1}^{n}\left\{y_i ln\left[\frac{\hat{\pi}(x_i)}{y_i}\right] + (1-y_i)ln\left[\frac{1-\hat{\pi}(x_i)}{1-y_i}\right]\right\}. \tag{5}$$

The statistic D often is called the deviance (it is also known as -2 *log likelihood* ($-2LL$)) and plays for the logistic regression the same role that the residual sum of squares plays in linear regression, it is a measure of how well the estimated model fits the data. The smaller the deviance is, the better the model fits the data.

To ascertain the significance of an independent variable it is necessary to compare the value of D with and without the independent variable in the equation, i.e.:

$$G = D(model\ without\ the\ variable) - D(model\ with\ the\ variable). \tag{6}$$

The statistics G plays the same role in logistic regression as the numerator of the partial F-test plays in linear regression and can be expressed by:

$$G = -2ln\left[\frac{likelihood\ without\ the\ variable}{likelihood\ with\ the\ variable}\right]. \tag{7}$$

Under the null hypothesis that the t "slope" coefficients for the covariates in the model are equal to zero, the distribution of G will be chi-square with t degrees-of-freedom.

The other similar, statistically equivalent test is the univariate Wald test. It is obtained by comparing the maximum likelihood estimate of slope parameter, $\hat{\beta}_i$, to an estimate of its standard error:

$$W_i = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)}. \tag{8}$$

Under the hypothesis, that individual $\hat{\beta}_i = 0$, the distribution of univariate Wald test statistics will be the Standard normal distribution. (Hosmer & Lemeshov, 2000)

2.1 Auxiliary variables

First of all the data were taken from the Population Register:

- Sex;
- Age;
- Citizenship;
- Nationality;
- Living region;
- Marital status;
- Country of birth.

Characteristics were recoded as dichotomous variables (i.e. is coded as either 0 or 1) for each individual, for example,

$$x_{1i} = \begin{cases} 1, & \text{if person is male} \\ 0, & \text{if person is female.} \end{cases}$$

Through in *SPSS* conducted experiment we could see, that *Nagelkerke R Square* value were lower than 0.5 and -2 *Log likelihood* statistics value is large, so we can conclude, that this information was not sufficient. Additional data was taken from other available sources of information:

- Information of employment;
- Whether the person is employed or self-employed;
- Personal income of the period (continuous variable).

2.2 The experiment and some conclusions

Using the logistic regression were computed the estimates of beta coefficients and probabilities of the case, that person is the resident of Latvia.

Through experiment we could see that the minimum probability of a case that a person is the resident of Latvia is quite large (about thirty percent). The estimated value of residents is much greater than the real value. Probably the main problem is that the number of emigrants in the investigated population part is too negligible (only 3% of the population). From the other side the unknown population part can be different from population part which the whole information is available. It was considered to select a balanced sample (balancing on the independent variables totals of unknown population part).

2.3 Balanced sampling

Definition. (Tillé, 2006) A sampling design $p(s)$ is said to be balanced with respect to the auxiliary variables $x_1, x_2, ..., x_p$, if and only if it satisfies the balancing equation given by

$$\hat{X}_{HT} = X. \tag{9}$$

which can also be written

$$\sum_{k \in U} \frac{x_{kj} S_k}{\pi_k} = \sum_{k \in U} x_{kj}, \tag{10}$$

for all $s \in S$ such that $p(s) > 0$ and for all $j = 1, ..., p$; or in other words $var(\hat{X}_{HT}) = 0$. Where s_k is the indicator whether the unit is in the sample or not, i.e.:

$$s_k = \begin{cases} 1, & \text{if unit } k \text{ is in the sample} \\ 0, & \text{if unit } k \text{ is not in the sample} \end{cases}$$

and $S_n = \left\{ s \in S | \sum_{k \in U} s_k = n \right\}$.

2.4 The experiment

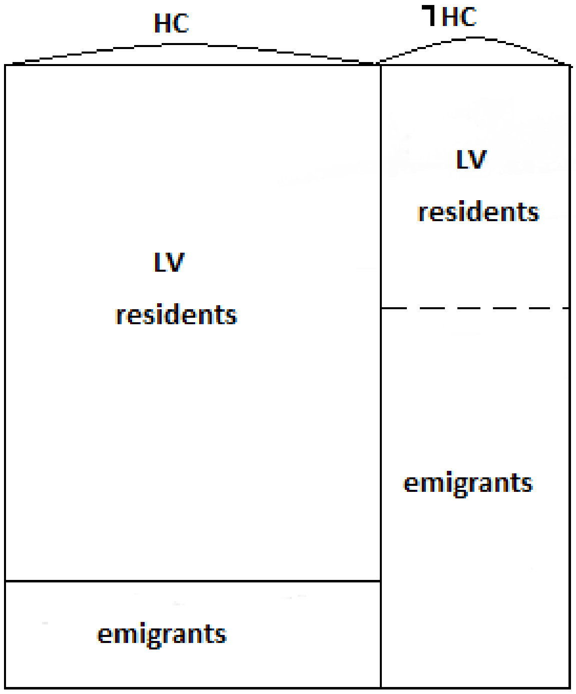
The aim was to select the sample which is balanced on the independent variables totals of unknown population part. The sample size were chosen directly proportional to unknown population part, i.e. $l = \frac{n_z}{N_n}$, and $N_n \leq n_z \leq N_z$, where N_z is the known population part size, N_n is unknown population part size and n_z is sample size. So the sample totals must be proportional to unknown population part totals, i.e.: $t_n = l * t_z$.

The obtained sample include only 5% emigrants and the result is the same as in previous experiments.

Conclusion

Through the experiments we have obtained the conditional probability $g_i = P(i \in LV | i \in HC)$, which means the probability, that person is the resident of Latvia, in condition that this person was participate in Housing Census 2011. To obtained the probability of unknown population part, we have to estimate the probability $p_i = P(i \in LV | i \notin HC)$. It will be the following step in this research.

Figure 1: Emigrants and residents of Latvia in Housing census 2011



References

Hosmer, D. & Lemeshov, S. (2000). *Applied logistic regression.* A Wiley Interscience Publication JOHN WILEY and SONS, ONC.

Tillé, Y. (2006). *Sampling algorithms.* Springer Sciens+Business Media, Inc.

CSB home page (Population Census): http://www.csb.gov.lv/en/statistikas-temas/population-census-30761.html

Business sample co-ordination

Jeļena Voronova[1]

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

Abstract

Most of the samples from a Business Register in Statistics of Latvia are independent from each other and basically not coordinated before. Response burden was the main purpose for starting business samples coordinating in 2011/2012. In fact, there are more than one positive aspect for statistics in final. Samples coordinating is based on Sweden method called SAMU and technique JALES. Method is based on permanent random number that is permanently associated with each Business Register unit. 12 samples with different types and purposes are placed on a one scheme for maximum negative co-ordination between them. No positive co-ordination was required from the head of surveys. In addition, the scheme was made with thoughts to increase and expand it, wherever it would be needed. So there are “places” for more surveys with negative and positive co-ordination opportunity with existing dozen. First real unit rotation of coordinated samples will be implemented in 2012/2013.

Keywords: Coordination, business statistics

1 Introduction

Response burden is actual problem in Business statistics in Latvia since time is money. Several methods are used, as decrease number of questionnaires and length, wide use of administrative data, sample size reducing, survey planning and so on. Main purpose is to find optimal and right way between statisticians eternal problem as estimation accuracy, respondent burden, survey planning, cost minimization...

This paper is concentrated on method of samples co-ordination in Latvia’s Business Statistics with aim to decrease respondent burden, without any important changing in surveys implementation. Business statistics samples in Latvia based on data from Business Register (BR) and implement as stratified simple random sample (SSRS). All surveys are independent (if there are no additional terms in query) from each other with personal frame, methods. Mathematical support division (MSD) receives samples queries for long-term (current and next years) and short-term (next year) statistics every year in October. As well as actual list of active enterprise from BR. All samples are ready to be used in early December and they will “work” all next year.

2 Sample co-ordination

2.1 Business register and samples

Sampling of all probability samples is done by MSD staff. Frame population used for calculating samples is based on the BR at Statistics Latvia. In November 2010 the actual sample frame contained 88313 enterprises.

There are several special rules and methods in the BR that let distinguish active unit from non-active. 22 samples queries for long- and short-term statistics were received at the end of 2010 (base data for analyzing). 18 samples were calculated in result (there are surveys with single aim and form, exceptions is length of form). Sub-samples and samples with survey period longer than a year were removed from a co-ordination list, and kept in mind. Only 12 samples could be coordinated in time. All enterprises were broken down into five groups by number of employees. In general, there is census in groups of "large" units (more than 50 employees), no sample and administrative data uses with "small" units (9 employees and less) and "middle" enterprises with employee from 10 to 49 employees surveys by random sample (Table 1). It means that only 10.36% of whole population could be coordinated into 12 samples during one time period.

Table 1: Population breakdown by groups of number of employees

Number of employee	Group	Number of enterprises	Proportion of enterprises, (%)
0-9	0	76677	86,82%
10-19	**1**	**5695**	**6,45%**
20-49	**2**	**3453**	**3,91%**
50-249	3	2090	2,37%
250+	4	398	0,45%
	Total	88313	100,00%

Based on 12 non-coordinated samples, which were calculated in November 2010, we can conclude that 8 is the maximum number of forms that an enterprise received to fill during 2011. The main aims of co-ordination is to decrease maximum number of forms what receive one enterprise and the number of enterprises, which do not receive any form from Central Statistics bureau of Latvia; and also to increase the number of enterprises which receive only 1 to 3 forms.

Table 2: Population distribution by number of forms

Number of survey forms	Number of enterprise	Proportion of enterprises, (%)
0	50139	56,8
1	10006	11,3
2	4712	5,3
3	2403	2,7
4	1254	1,4
5	1025	1,2
6	644	0,7
7	54	0,1
8	1	0,0
Total survey:	70238	79,5
None:	18075	20,5
Total:	88313	100

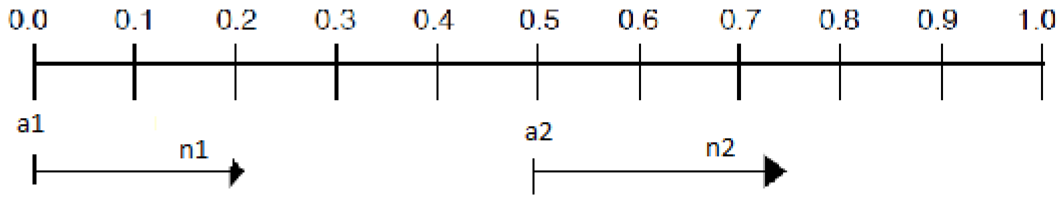
2.2 Co-ordination methods

The experience of other countries in sample coordinating was studied, to detect methods and ways that are closer for existing sampling methods in Latvia.

Swedish method (SAMU/JALES)

The technique used by the SAMU to select samples was developed at Statistics Sweden in the late 60's. SAMU generates samples from updated BR. Samples are drawn by so called JALES technique is based on permanent, independent and unique random number (PRN),which is uniformly distributed over the interval (0,1) and permanently associated with every BR unit. PRN birth and death occurs at the same time as its unit birth and death. Each sample is calculated using only PRN. To select simple random sampling, we can take *n* units from any point in the interval (0,1) and any direction from this point. If there is not enough unit at the end (start) of interval, we can continue selection from the start (end) of it. To co-ordinate two samples with sample size n_1 and n_2 it's only necessary to detect starting point a_1 and a_2 in the interval (0,1) and direction (left or right) to start selecting unit (Figure 1).

Figure 1: Two samples co-ordination



We have to choose the same starting point and direction for both samples to get the maximum positive co-ordination. Select different well apart starting point and use the same directions for negative co-ordination of two samples. If there are not enough units to provide complete negative co-ordination, technique could reduce overlapping.

Unit rotation is needed in SAMU due to the positive co-ordination in time. The last digits in PRN are associated unit to random group. There are five rotation groups in SAMU. One group is rotating in one year by shifting random number (or starting point) on 0.01

Netherland method (EDS)

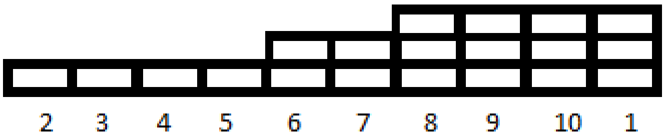
There are used PRN to co-ordinate samples in Netherland. Method based on burden measurement by estimated time required for complete the form, dividing frame into six classes (Table 3).

Table 3: Class of response burden in EDS

Class	Completion time (min)	Response burden
1	1-30	1
2	31-60	2
3	61-120	4
4	121-180	6
5	181-240	8
6	241+	10

After every sample calculation completion time is accumulated for every unit in the frame and set response class. Before every next sample selection, units are sorted by the PRN and response class (ascending). Thus the most burdened units are sent at the end of the list (Figure 2).

Figure 2: Coordinating samples in EDS after two selections



The file with unit's identification numbers and dummy variables that indicates unit being in the sample for last period was generated for unit rotation. Rotation fraction is calculated taking in account total number of enterprise and possibility to remove unit from next year survey sample. Rotation fraction is in the interval [0;1], where 0 means that next year sample will be maximum positive coordinated with last year sample.

Other methods

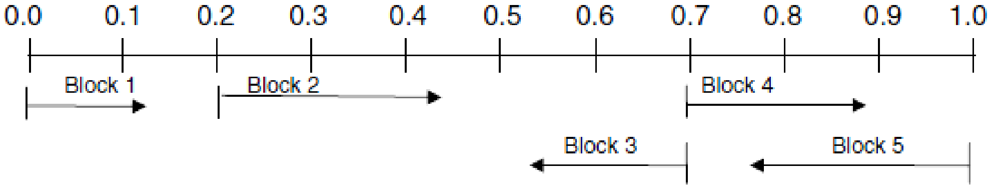
France coordination method (OCEAN) was also investigated. It is based on assigning random number, which are recalculated after each sample. Sample selection and unit random methods are similar to Swedish SAMU. Finland method OTKO is relatively young. There is used PRN in this method and response index, equal to 100 at the start. Response index decrease after each unit selection in the sample. When index reduces till determinates level, unit comes out of frame on a time. Then it came back with started index 100.

SAMU method with JALES technique could be used in Statistics Latvia, without global changes in sample design, as well as easy understandable administration and suitable to main guidelines of business survey sample.

2.2 Co-ordination method in statistics Latvia

Optimally best coordinated scheme need to be constructed for Statistics Latvia. All 12 samples should be divided on blocks. Five blocks (Figure 3) is used in SAMU. Left sample direction and positive co-ordination within one year would not use in statistics Latvia.

Figure 3: Five block in the SAMU by Background facts on Economic statistics *et al.* (2003, p.15)



Its own starting point *a* was detected first for every survey in the interval (0,1). After getting results, surveys were grouped and changing its location in several occasions. In practice from twelve samples needed in co-ordination, 7 surveys studies different economics activities (NACE rev2.) and mostly have no overlapping. And there are 5 surveys which target population is mostly whole population of active enterprise (Table 4).

Table 4: Distribution Survey interest by 5 surveys

Number of survey	Number of enterprises	Proportion of enterprises, (%)
0	1	0.0
1	351	3.8
2	431	4.7
3	165	1.8
4	1139	12.5
5	7061	77.2
Total	9148	100.0

Keep in mind only enterprises, with 10 to 49 employees could be coordinated. An optimal scheme (Table 5) was created after swapping block and analyzing result. Best location for every survey was chosen, taking in account:

- Number of questionnaires, that enterprise will receive in time;
- Estimated time required to fill questionnaire.

Each survey got its own place and starting point at co-ordination scheme.

Table 5: Co-ordination scheme for Business Statistics in Latvia since 2011/2012

Block number	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Surveys	7 studies with no overlapping population frafe	1 survey	1 survey	Free place for rare and special surveys	1 survey	1 survey	1 survey
Starting point	**0**	**0.143**	**0.286**	**0.429**	**0.571**	**0.714**	**0.857**

Permanent random number is calculated and fixed with rounding 9 digits after decimal point. Unit rotation, after BR advice, will occur in 3 year period. Frame is divided into 3 groups, by the last three digits in PRN. First rotation into frame will occur in November 2012.

Main achievements we have:

- Negative co-ordination between surveys (it is impossible to get entirely non-overlapping samples in situation with small population);
- Positive co-ordination in time;
- Response burden.

References

Background facts on Economic Statistic (2003). SAMU. The system for co-ordination of frame populations and samples from the Business Register at Statistics Sweden. *Statistics Sweden, 2003:3*

Coordination of samples: the microstrata methodology. SALOMON. Eurostat project.

Mészáros, P. A program for sample co-ordination SALOMON, Contributed paper

Nedyalkova, D., Pea, J., Till lé, Y.(2009). A Review of Some Current Methods of Coordination of Stratifed Samples. Introduction and Comparison of New Methods Based on Microstrata

Ohlson, E. (1992) SAMU. The system for Co-ordination of samples from the Business Register at Statistics Sweden. *Statistics Sweden , R&D report 1992:18*

Teikari, I. (2001). Poisson mixture sampling in controlling the distribution of response burden in longitudinal and cross section business survey. *Research reporst* 232, Statistics Finland.

Valliant, R. (2007) Survey sample coordination. A Summer School co-arranged by Orebro University and Statistics Sweden. *University of Michigan, U.S., Joint program in Survey Methodology University of Maryland.*

A Simulation Study on Nonresponse-bias for Calibration Estimator with Missing Auxiliary Informaiton

Lisha Wang

Örebro University, e-mail: lisha.wang@oru.se

Abstract

The calibration approach is suggested in the literature for estimation in sample surveys under non-response given access to suitable auxiliary information. However, missing values in auxiliary information come up as a thorny but realistic problem. This paper is connected with how imputation of auxiliary information based on different levels of register information affects the calibration estimator. Results show that the level of register information used for deriving imputation models only marginally affects the calibration estimator bias. The results are obtained under different patterns of non-response in the target variable and missing values of the auxiliary variable.

Keywords: Nonresponse, calibration, imputation, bias, auxiliary variable

1 Introduction

The calibration approach is by Särndal & Lundström (2005) suggested for estimation in sample surveys with non-response. Calibration implies computation of weights for sampled elements in the response set such that applied to known auxiliary variables they replicate known population totals. There are several papers addressing the calibration technique for estimation in sample surveys. Deville & Särndal (1992) proposed linear form for the calibration weighting with multivariate auxiliary information. Kott (2006) considered calibration estimation to correct for coverage errors and unit non-response (Quasi-randomization). Montanari (2005) discussed calibration estimator in a neural network mode. Särndal & Lundström (2008) discussed non-response bias for choosing auxiliary information.

In these papers, it is usually assumed that auxiliary variables are fully recorded without missing values. For example, Särndal & Lundström (2005) proposed star vector and moon vector, which are defined as information available at the population level and the sample level, respectively. Unfortunately, in reality, auxiliary information is not that ideal. Missing values occur in all types of data, also in auxiliary variables records.

Missing values can be replaced by imputations and then treated as any other auxiliary variable. However, this depends on the way the imputations are derived. The cases when missing values are replace by a constant, zero say, and the case when imputations are derived from a regression model estimated on response set information are different. This paper addresses the issue of the effects of how imputations of auxiliary variables are derived. Using simulation, the bias of the calibration estimator is studied under regression imputation where the regression imputation model is estimated using register, sample, and response set information respectively.

2 Calibration Estimator

2.1 Definition

Consider a finite population with N elements U=$1, 2, \ldots, N$, in which y_k is a target variable and $x_k = (x_{1k}, x_{2k}, \ldots, x_{Jk})'$ is a full-recorded auxiliary vector. A probability sample s with sample size n is selected from U by a probability sampling design $p(s)$. When non-response occurs, only a subset of the sample $r \in s$ is answered, where the size of response set is denoted as n_r. To describe the random response mechanism, $q(r|s)$ is denoted as the conditional response distribution, and the probability of a response of element k given its selection to a sample is denoted $\theta_k = Pr(k \in r|k \in s)$.

To estimate the population total $Y = \sum_U y_k$, the calibration estimator $\hat{Y}_w = \sum_r w_k y_k$ uses calibrated weights w_k subject to the constraint $\sum_r w_k x_k = X$. The weights w_k can be defined in different ways obeying the constraint. Särndal & Lundström (2005) defined the weights using the system $w_k = d_k v_k$, $v_k = 1 + \lambda_r \mathbf{x}_k$, and $\lambda_r = (X - \sum_r d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}$.

2.2 Auxiliary information

Särndal & Lundström (2005) defined three different cases depending on the accessible auxiliary information. In this paper, we will look into two of them. The two cases are

InfoU. Information is available at the level of the population U such that

- the population total $\sum_U \mathbf{x}_k^\star$ is known;
- for every $k \in r$, the value of $\mathbf{x}_k^\star$ is known.

InfoS. Information is available at the level of the sample s such that

- for every $k \in s$, the value of $\mathbf{x}_k^\circ$ is known but $\sum_U \mathbf{x}_k^\circ$ is unknown.

Consider the case that missing values occur in auxiliary variable x_k as well. Imputation is a frequently-used method to allocate artificial values for the missing items. Little & Rubin (1987) regarded imputation as a general and flexible method for handling missing-data problem but with pitfall, such as substantial bias, and summarized different sorts of imputation methods to construct the substitutes.

With imputed values, auxiliary variable will be denoted as

$$x_{\bullet k} = \begin{cases} x_k & \text{for } k \in r_x \\ x_k(\hat{\delta}) & \text{for } k \in U - r_x \end{cases} \tag{1}$$

here r_x is the subset of the population U where x_k is available, and $x_k(\hat{\delta})$ is the imputed value based on the parameter $\hat{\delta}$ which is derived from the register information.

Consider InfoU with $\mathbf{x}_k = \mathbf{x}_k^\star = (1, x_{\bullet k})'$ with the information input $\mathbf{X}_{\bullet k} = (N, \sum_U x_{\bullet k})$, where the calibration estimator for target variable y becomes

$$\hat{Y}_w = N\bar{y}_r + (\sum_U x_{\bullet k} - N\bar{x}_r) * B_r$$

where

$$B_r = \frac{\sum_r d_k (x_{\bullet k} - \bar{x}_r)(y_k - \bar{y}_r)}{\sum_r d_k (x_{\bullet k} - \bar{x}_r)^2} \tag{2}$$

$$\bar{y}_r = \frac{\sum_r d_k y_k}{\sum_r d_k} \tag{3}$$

$$\bar{x}_r = \frac{\sum_r d_k x_{\bullet k}}{\sum_r d_k} \tag{4}$$

Consider InfoS with $\mathbf{x}_k = \mathbf{x}_k^\circ = (1, x_{\bullet k})'$ with the information input $\mathbf{X}_{\bullet k} = (\hat{N}, \sum_s d_k x_{\bullet k})$, where $\hat{N} = \sum_s d_k$. The calibration estimator for target variable y then becomes

$$\hat{Y}_w = \hat{N}\bar{y}_r + (\sum_s d_k x_{\bullet k} - \hat{N}\bar{x}_r) * B_r$$

where B_r, $\bar{y}_r$ and $\bar{x}_r$ are the same as in equation (2), (3) and (4).

2.3 Nearbias

A central issue regarding the effects of nonresponse is estimation bias. Consider an auxiliary vector x_k satisfying $\mu'\mathbf{x}_{\bullet k} = 1$ for all k. Then Särndal & Lundström (2005) shows

$$Nearbias(\hat{Y}_w) = (\sum_U \mathbf{x}_{\bullet k})'(\mathbf{B}_{U;\theta} - \mathbf{B}_U) \tag{5}$$

in which

$$\mathbf{B}_{U;\theta} = (\sum_U \theta_k \mathbf{x}_{\bullet k}\mathbf{x}'_{\bullet k})^{-1}(\sum_U \theta_k \mathbf{x}_{\bullet k} y_k)$$

and

$$\mathbf{B}_U = (\sum_U \mathbf{x}_{\bullet k}\mathbf{x}'_{\bullet k})^{-1}(\sum_U \mathbf{x}_{\bullet k} y_k)$$

Consider the case that missing value occurs in auxiliary variable x_k, x_k becomes $x_{\bullet k}$ as described in equation (1). If all the missing items in x_k is artificially imputed as 0, i.e., $x_{\bullet k} = 0$ when $k \in U - r_x$, the formula of nearbias will be the same as equation (5) with no extra components.

If we consider another case in which regression imputation is utilized for missing values in x_k, i.e., x_k is replaced by $x_{\bullet k}$ and $x_{\bullet k} = x_k(\hat{\delta})$ when $k \in U - r_x$, where $\hat{\delta}$ is the estimate of coefficient derived from the register system. The formula of nearbias in this case will stay the same as equation (5).

According to accessible register information at hand, the estimate of $\hat{\delta}$ could possibly derived from population level, response level or sample level. When $\hat{\delta}$ is obtained from register information only, equation (5) will keep valid through all the cases.

In next section, a simulation study will be conducted on how bias of calibration estimator changes when different imputation for missing values in x_k is used. Missing values in x_k will be imputed by linear regression model $x_k = \mathbf{u}'_k\boldsymbol{\delta} + \varepsilon_k$ and $\boldsymbol{\delta}$ is estimated based on sample level or response level.

3 Simulation Study

The effect of imputation on the calibration estimator bias is here studied by simulation. To simulate a population with 100000 units, the following procedures are performed.

1. x_k is generated from a standard normal distribution N(0,1).
2. error term ξ_1 and ξ_2 are independently generated from N(0,1) distribution.
3. u_k is generated by $u_k = \alpha + \beta * x_k + \rho_1 * \xi_{1k}$.
4. y_k is generated by $y_k = \tau + \eta * x_k + \rho_2 * \xi_{2k}$.

The coefficients β, η, ρ_1 and ρ_2 are used to control the coefficient of determination R^2 between y and x, and x and u respectively.

The bias of the calibration estimator $Bias(\hat{Y}_w) = E(\hat{Y}_w) - Y$ will be studied in four different cases with different patterns of the occurance of non-response in y_k and missing values of x_k.

Case I non-response in y_k occurs with constant probability such that $\theta_k = \theta$ for all $k \in U$, and missing vaule of x_k occurs with constant probability such that Pr(x_k is missing in register system)=ϑ_k=ϑ for all $k \in U$.

Case II missing vaule of x_k occurs with constant probability such that ϑ_k=ϑ for all $k \in U$, but non-response in y_k occurs with varying probability, i.e., θ_k changes for $k \in U$.

Case III non-response in y_k occurs with constant probability such that $\theta_k = \theta$ for all $k \in U$, but missing vaule of x_k occurs with varying probability, i.e., ϑ_k changes for all $k \in U$.

Case IV non-response in y_k and missing value in x_k both occur with varying probability, i.e., both θ_k and ϑ_k change for all $k \in U$.

In Case II and IV above, y_k is divided into three groups, with response rate

$$\theta_k = \begin{cases} 10\% & \text{when } y > 8 \\ 35\% & \text{when } y < 0 \\ 65\% & \text{when } 0 \leq y \leq 8 \end{cases}$$

Similarly, in Case III and IV, x_k is divided into ten groups, with response rate

$$\vartheta_k = \begin{cases} 45\% & \text{when } x < -1.28 \\ 50\% & \text{when } -1.28 \leq x < -0.84 \\ 55\% & \text{when } -0.84 \leq x < -0.52 \\ 60\% & \text{when } -0.52 \leq x < -0.25 \\ 65\% & \text{when } -0.25 \leq x < 0 \\ 75\% & \text{when } 0 \leq x < 0.25 \\ 80\% & \text{when } 0.25 \leq x < 0.52 \\ 85\% & \text{when } 0.52 \leq x < 0.84 \\ 90\% & \text{when } 0.84 \leq x < 1.28 \\ 95\% & \text{when } x \geq 1.28 \end{cases}$$

The regression imputation will be utilized to make up for the missing values in auxiliary variable x_k. In this stage, imputation will be proceeded based on the linear model

$$x_k = \mathbf{u}_k'\boldsymbol{\delta} + \varepsilon_k = \delta_1 + \delta_2 u_k + \varepsilon_k$$

where u is a full-recorded variable from the register system. The estimator of $\boldsymbol{\delta}$ is $\hat{\boldsymbol{\delta}} = (\sum_A \mathbf{u}_k\mathbf{u}_k')^{-1}(\sum_A \mathbf{u}_k x_k)$, where A is the set of items used for performing the linear regression. The following three kinds of collection of objects (i.e., A) are considered.

Imputation 1 $A = U_x$, where U_x is the whole population of variable x, which means imputation regression will be run based on all the available values of x_k in the population.

Imputation 2 $A = r_x = U_x \cap r$, where imputation regression will be executed based on the available values of x_k in the population where y_k is responsed.

Imputation 3 $A = s_x = U_x \cap s$, where imputation regression will be performed based on the available values of x_k in the sample.

Replicating the simulation for 3000 times, the expectation of the calibration estimator is estimated by $E(\hat{Y}_w) = \sum_{i=1}^{3000} \hat{Y}_{w_i}/3000$ and the bias is estimated with $Bias(\hat{Y}_w) = E(\hat{Y}_w) - Y$. As a benchmark, estimates of the bias of the calibration estimator with full-recorded auxiliary variable x_k is shown in Table 1. The bias in the case II/IV is notably larger than in case I/III.

Bias estimates for the calibration estimator with missing values of the auxiliary variable are reported in tables 2 - 5. In Table 2, the case with $R^2(y, x) = R^2(x, u) = 50\%$ is considered. From the table it is seen that the bias estimates are small for cases I and III, while they are large in cases II and IV. A surprising observation is that bias estimates are in large unaffected by the level of information used for estimation of the regression model used for imputation. Also, bias is in large unaffected by the level of information used in the calibration estimator. However, compared to the beachmark in Table 1, the biases in tables 2 - 5 are larger in general.

In the following tables, bias estimates are reported for different cases of strengths in the relation between y and x, and between x and u. Compared with Table 2, Table 3 reports results where $R^2(y,x)$ is increased to 0.8. It is observed that the results in the two tables are comparable. There are only minor changes in the bias estimates.

In Table 4, bias estimates in the case when $R^2(x,u)$ is increased to 0.8, compared with Table 2, is reported. Again, the results are comparable with those of Table 1 with only small difference in bias estimates. Finally, Table 5 reports results when $R^2(x,u)$ is decreased to 0.26 and it is seen that the reported estimates are comparable with those of tables 2 - 4.

Table 6 reports a case when the auxiliary variable x_k is chi-square distributed instead of normal as in tables 1 - 5. The levels of the bias estimates are different (see Table 6) but the general pattern observed in tables 2 - 5 is also observed here. Biases are negligible in cases I and III, while modest in cases II and IV. The level of the bias is not dependent of the level of information used in the calibration, and finally, the bias is largely unaffected by the level of information used for the estimation of the regression relation used as imputation model.

4 Conclusion and Discussion

It is shown in all the cases of the simulation study that the bias of calibration estimator differs slightly between InfoU case and InofS case, which is also stated in Särndal & Lundström (2005). The bias estimates in Case I/III are always close to 0, implying that the calibration estimator is nearly unbiased when the response rate θ_k is constant and not related to the value of y_k. The bias estimates in Case II/IV, however, are affected by θ_k being related to the value of y_k.

The simulation study also shows that imputation of auxiliary information with coefficient $\hat{\delta}$ derived from different levels of information, i.e., register, sample and response set give negligible differences on bias estimates. This implies that when missing values in auxiliary information need to be imputed, the level of information used for imputation has no essential effect on the bias of the calibration estimator. The results are here derived by simulation and it is desirable to derive more formal and general results. One suggestion for further studies is to consider the asymptotic properties of Y_w, where asymptotic limits $f(\hat{\delta})$ are utilized. It is expected that bias expressions similar to equation (5) can be derived using asymptotics.

References

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

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

Little, R. & Rubin, D. (1987). *Statistical analysis with missing data.* John Wiley & Sons.

Montanari, M., G.E. & Ranalli (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association* **100**, 1429 – 1442.

Särndal, C. & Lundström, S. (2005). *Estimation in surveys with nonresponse.* John Wiley & Sons.

Särndal, C. & Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics* **24**, 167 – 191.

A Table: simulation result

Table 1: Bias in Normal case when $R^2(y,x)$=$R^2(x,u)$=50%

	Case I	Case II
InfoU	139.28	-11884.92
InfoS	350.78	-11616.12

Note: x_k is full-recorded.
$\sum_U y_k$=500915.62

Table 2: Bias in Normal case when $R^2(y,x)$=$R^2(x,u)$=50%

		Case I	Case II	Case III	Case IV
InfoU	Imputation 1	-100.82	-13307.32	167.32	-11509.29
	Imputation 2	-105.22	-13345.59	137.73	-11275.61
	Imputation 3	-102.62	-13343.70	128.98	-11547.56
InfoS	Imputation 1	76.68	-13074.37	455.78	-11223.00
	Imputation 2	84.08	-13091.86	385.32	-10943.10
	Imputation 3	19.18	-13078.22	457.31	-11271.17

Note: $\sum_U y_k$=500915.62

Table 3: Bias in Normal case when $R^2(y,x)$=80% and $R^2(x,u)$=50%

		Case I	Case II	Case III	Case IV
InfoU	Imputation 1	-170.51	-13714.31	97.41	-10606.63
	Imputation 2	-179.55	-13556.80	67.67	-9100.94
	Imputation 3	-158.29	-13693.62	82.72	-10590.23
InfoS	Imputation 1	77.15	-13466.74	378.16	-10386.30
	Imputation 2	37.92	-13290.21	326.84	-8877.21
	Imputation 3	49.62	-13504.88	332.76	-10401.46

Note: $\sum_U y_k$=500967.42

Table 4: Bias in Normal Case when $R^2(y,x)$=50% and $R^2(x,u)$=80%

		Case I	Case II	Case III	Case IV
	Imputation 1	16.13	-12462.12	109.29	-11716.29
InfoU	Imputation 2	18.85	-12490.73	100.82	-11596.08
	Imputation 3	-7.76	-12504.85	65.81	-11755.04
	Imputation 1	275.90	-12216.29	382.02	-11438.96
InfoS	Imputation 2	193.84	-12231.55	296.71	-11291.59
	Imputation 3	246.78	-12257.83	338.56	-11483.31

Note: $\sum_U y_k$=500915.62

Table 5: Bias in Normal Case when $R^2(y,x)$=50% and $R^2(x,u)$=26%

		Case I	Case II	Case III	Case IV
	Imputation 1	-173.81	-13913.94	231.92	-11403.91
InfoU	Imputation 2	-177.67	-13947.07	196.40	-11148.80
	Imputation 3	-183.65	-13953.12	200.92	-11432.49
	Imputation 1	77.05	-13698.89	522.33	-11111.83
InfoS	Imputation 2	0.48	-13705.81	443.75	-10801.25
	Imputation 3	41.84	-13743.01	478.42	-11165.83

Note: $\sum_U y_k$=500915.62

Table 6: Bias in chi-square case when$R^2(y,x)$=$R^2(x,u)$=85%

		Case I	Case II	Case III	Case IV
	Imputation 1	-1575	-21848	-1408	-22164
InfoU	Imputation 2	-1649	-21534	-1484	-20986
	Imputation 3	-1668	-21898	-1527	-22184
	Imputation 1	-1171	-21368	-1012	-21743
InfoS	Imputation 2	-1078	-20908	-920	-20406
	Imputation 3	-1027	-21325	-925	-21668

Note: $\sum_U y_k$=1099883.12

Current Development in Microsimulation and Assessment of Uncertainty in JUTTA Model

Meng Zhou[1,2] and Maria Valaste[2]

[1]University of Helsinki, email: meng.zhou@helsinki.fi
[2]Finnish Social Insurance Institution

Abstract

Nowadays, microsimulation method has been introduced to different fields, such as Social Science, Medicine research and Economic study. This method evaluates the effects of the proposed interventions or policies before they are implemented in the real world. In this article, we will concentrate on microsimulation method used in Social Science by firstly explaining two main streams in microsimulation world, Static approach and Dynamic approach. In the following section, the uncertainty of a Finnish static microsimulation model JUTTA is assessed and Toimtuki model one of the sub-model in JUTTA is detected to have space to be more accurate. In order to do so, two statistical models- Linear Regression model and Two-Stage Least Squares (2SLS) model are applied to it. From the results, we could conclude that both the Linear Regression and 2SLS successfully improves the accuracy of TOIMTUKI to some extent.

Keywords: Static microsimulation, Dynamic microsimulation, JUTTA, assessment, 2SLS

1 Introduction

What and Why

A microsimulation model differs from other types of models in that it operates on individual units rather than on aggregate information (TRIM3. 2012a). Typically, in social sciences, those units are individual substantial or economic units. The database used as input to a microsimulation model contains records describing persons, households or business. And the simulation model applies a set of rules to each individual record. The result of the computations might be the amount of taxes owed by the unit to which the unit is entitled under certain government legislation. Also, if we are interested in the total tax, each individual result should be multiplied by whatever weight is associated with the unit in the microdata file, then the weighted individual results are added together to obtain the aggregate result. Thus, different policies could apply to the same microdata file, and the report of the comparisons among the different results could be a good helper to the wise government.

The purpose of the microsimulation is mainly to evaluate the effects of the proposed interventions or policies before they are implemented in the real world. By using microsimulation, people can easily estimate the impacts of a new scheme by producing outputs on a wide range of measures of effectiveness.

Currently, there are two main streams in the microsimulation field, which are Static microsimulation and Dynamic microsimulation.

2 State of the art

Static microsimulation

The Static microsimulation has one important character that it does not take the individual behavior into account, which means once the rules are made, they will be obeyed 100% without any variation. It suits for performing detailed simulations of past, the present, and the near future. It typically use static aging techniques, changing certain variables on the original microdata file to produce a file with the demographic and economic characteristics expected in the future year. Person weights are modified to change the total population and the weighted characteristics of the population; labor force status may be changed to alter the unemployment rate; and incomes are adjusted for price changes. Simulations can then be run on the aged microdata file to estimate the impact of a change to be implemented in the future year (TRIM3. 2012b).

Dynamic microsimulation

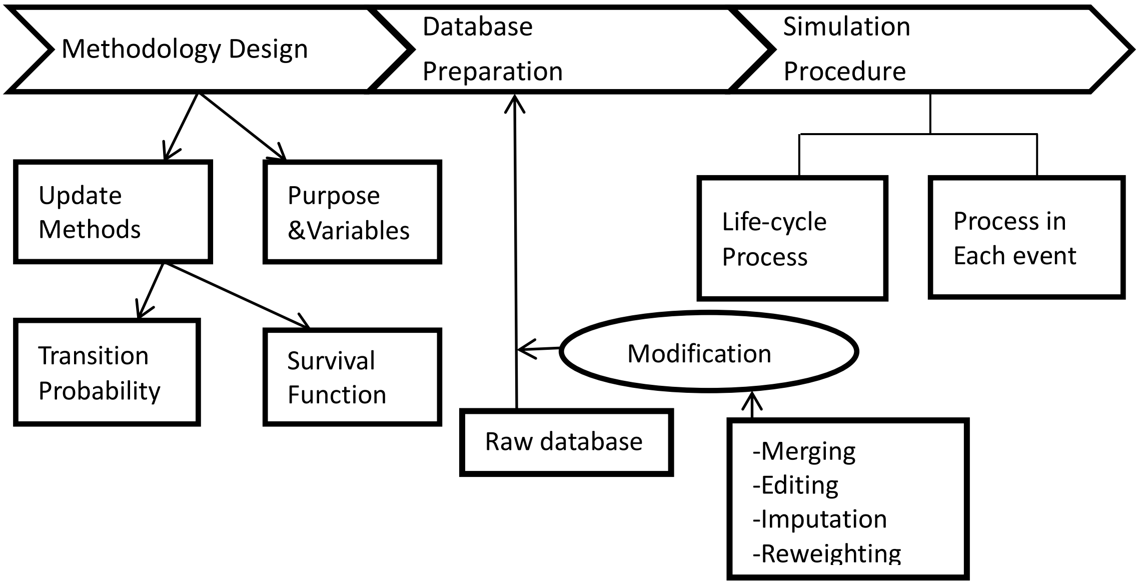
Dynamic microsimulation models age each person in the microdata file from one year to the next by probabilistically deciding whether or not that person will get married, get divorced, have a child, drop out of school, get a job, change jobs, become unemployed, retire, or die, then the same procedure is repeated as many times as the user wants to achieve the final simulation year. Simulations of government legislations can be run in the current year, the final year of the aging process, or any interim year. The simulation of the government program in one year may affect a person’s characteristics in the subsequent year (TRIM3. 2012b). For example, whether or not someone will drop out of school could be programmed to depend partly on family income, which could, in turn, be affected by government transfer payments. This kind of models could create the synthetic database for a future year, which is capable of performing simulations into the distant future, but it couldn’t capture as much details as static models do.

In Dynamic microsimulation models the transition probabilities play the important role, because they are used to create the synthetic database about the individuals’ life paths on the demographic events, personal events and so on.

3 The detailed Dynamic microsimulation

There are three components in the Dynamic Microsimulation: methodology design, database preparation and simulation procedure. The figure 1 presents the basic structure of it.

Figure 1: Simulation Procedure



Updating methods and statistical model application

In the methodology step, deciding the update methods is the main point. Usually, there are two options, transition probability and survival function. Nowadays, the most popular updating method is based on the transition probabilities, and the space in time between the updating processes is one year. Thus, the estimation of the transition probability becomes the hot point, where statistic models are mainly applied. The often used models are linear model, generalized linear model, mixed model and so on. The survival function method is sometimes used, for instance, the cox model is used when estimating the event fertility, please see the example in Anthony *et al.* (1999).

Here, we will talk about the application of Generalized Linear Model (GLM). The generalized linear model is a flexible generalization of ordinary linear regression. The linear model can be transformed to a generalized linear model by linked function g(). The model could be represented as: $E(Y) = \mu = g^{-1}(X\beta)$ Where $E(Y)$ is the expected value of *Y*, $X\beta$ is the linear predictor, a linear combination of unknown parameter vector β, g is the link function.

There are many commonly used link functions. In Dynamic microsimulation, the logit and probit are the two most useful models when we are estimating the transition probabilities. Here one example will be given, the event is the "Employment Status", for more information, please see Lennart Flood *et al.* (2005).

It would be a good case to illustrate the Monte Carlo Simulation. Monte Carlo technic gives the model stochastic property. For the binary variable employment status, we have a Bernoulli distribution, i.e. $Y_i \sim bernoulli(\pi_i)$, where $\Pr[Y_i = 1] = \pi_i$ and $\Pr[Y_i = 0] = 1 - \pi_i$. As an illustration, let Y_i denote unemployment for individual i during the period of interest. Let Y_i=1 denote unemployment and Y_i=0 denote employment, π_i denote the probability that the individual is unemployed during the year. This event is simulated by comparing π_i with a uniform random number u_i . If $u_i < \pi_i$ the event is realized and individual i become unemployed.

The propensity of becoming unemployed is determined by π_i , by allowing π_i to be determined by individual or household attributes these attributes also determine the probability of unemployment. This is typically accomplished by a logit regression model. The logit model is given as $\pi_i = [1 + \exp(-X_i\beta)]^{-1}$, where X_i is a vector of individual or household characteristics like gender, age, working history or any other characteristic relevant for explaining unemployment, i.e. rate of regional unemployment and β is a vector of parameters.

In order to calculate the estimator $\hat{\pi}_i$, firstly, we need to know the expected parameter vector $\hat{\beta}$, where it could be estimated from the outsource databases, such as registered database and official survey database. Then, the dependent variable $\hat{\pi}_i$ is calculated this way:

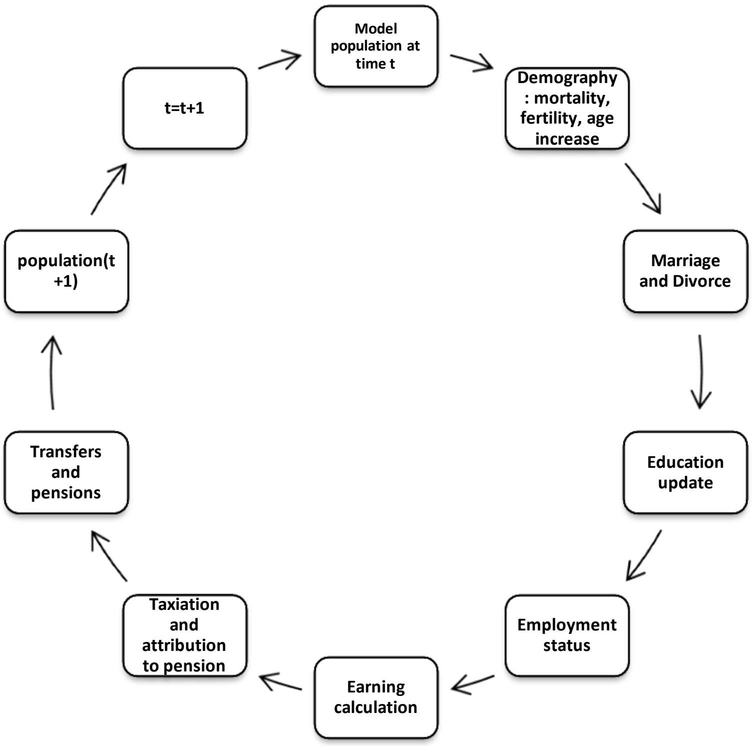
$$\hat{\pi}_i = [1 + \exp(-X_i\hat{\beta})]^{-1}$$

After obtaining $\hat{\pi}_i$, u_i is chosen randomly from the uniform distribution: $u_i \sim U(0,1)$. Finally simulated binary variable employment status is assigned to 1 or 0 by comparing $\hat{\pi}_i$ and u_i.

Simulation procedure: Model structure

The dynamic microsimulation model ages the underlying data base by one year, and that is run repeatedly to generate the multi-year demographic evolution needed for the whole simulation. Figure 2 describes us the life-cycle structure in the normal dynamic microsimulation model.

Figure 2: Life-cycle events process



Its "kernel" ages an input database by one year in any given pass. During each such pass, it simulates all of the births, deaths, marriages, labor force entry and exit and earnings, etc., that occur during that simulation year, and ages each of the individuals in the database by one year. It then outputs another database that is itself a new, representative population, but one that reflects the situation one year later than did the previous input database. The cycle is repeated over and over again for the length of the simulation run; in each cycle, the output data base from one pass through the kernel is used as the input for the next pass. (Anthony *et al.* 1999)

We could consider the aging process as a sequence of modules (events) that this step consists of a number of modules executed in(operated in) sequence, each of them modifying the in-memory population for that module's event for the current year. Each module processes all of the population for which that module/event is relevant, updating that aspect of the individuals' lives. However, not all individuals are eligible for all modules; e.g. individuals who have previously died will not give birth, and individuals who are presently married are not, in the same year, eligible to enter the marriage market (Rick Morrison, 1998).

Once the full set of modules has been executed, they have collectively aged the in memory population by one year. That is, they have implemented all of the events that effectively transform the base from one year's representative population to the next year's representative population.

4 Assessment of uncertainty of the JUTTA model and innovation method in TOIMTUKI

Background of JUTTA model

The JUTTA model is a static microsimulation model developed by Social Insurance Institution of Finland, it is also called tax-benefit model. JUTTA has 10989 households and around 30000 individuals sample size, and the data resources came from Statistics Finland. It has ten sub-models and one main model. The sub-models are designed for each branch of legislations and the main model is designed for running all the sub-models and producing the final results of the key data based on household level. The sub-models include: SAIRVAK, TTURVA, KOTIHUKI, OPINTUKI, KANSEL, VERO, LLISA, ELASUMTUKI, ASUMTUKI, TOIMTUKI. They represent sickness insurance benefits, unemployment benefits, child care benefits and day-care fees, study grant, the national pension system, personal taxes, benefits for families with children, pensioner's housing allowances, general housing allowances, means-tested income support, respectively. For each of these sub-models, parameter system and function system were built. (Honkanen Pertti, 2010)

Assessment of JUTTA model

In all the models, the accuracy is calculated in two different forms, one is the absolute difference percentage and the other one is the relative difference percentage. The absolute difference percentage is calculated based on the classifying the absolute errors between the real value and estimated value to the intervals [0, 1), [1, 10), [10, 100), [100, 1000) and [1000, ∞), and then dividing the number of the observations in each intervals by the total number of the observations. The relative difference percentage is similar to the absolute one, but classifying the errors in the intervals [0%, 0.1%), [0.1%, 1%), [1%, 10%), [10%, ∞).

After obtaining the percentage results, we could see that most of the models perform quite well, with their variables' accuracy high enough in the interval (60%, 100%] for both absolute different and relative different in first level called [0, 1) and [0, 0.1%) respectively. However, there is one extremely inaccurate model called TOIMTUKI meaning income-related supplementary benefit, which with both zero percentage in the first level intervals and more than 60% in the last intervals([1000, ∞) and [10%, ∞)).

The Toimtuki calculates the last benefit the people could apply after house benefit, heath benefit, student benefit and so on. In the other word, the Toimtuki could be regarded as the "residual" benefit in the JUTTA model, where the people apply when no other benefits could be applied.

In order to improve Toimtuki's accuracy, the statistical models are used. The first method is the Linear Regression method. After checking the relevant variables, seven of them are significant, which are tyot, ttyotpr, svatva, maksvuok, jasenia, desmod, meaning number of month of person's unemployment or forced leaving, unemployment allowance, household yearly income, monthly house rent, number of people in the household, decile group the household belonging to (according to OECD).

Method one: The linear regression model

$$\hat{y} = X\hat{\beta} \quad (1)$$

where X is the characteristics mentioned vector, $\hat{\beta}$ is the vector of estimated coefficients and $\hat{y}$ is the estimated benefits the household should receive. Now we plug the numeric coefficients in to equation (1):

$\hat{y}$ =195.67209+158.40442*tyot-0.17366*ttyotpr-0.10823*svatva+2.02230*maksvuok

+1308.86787*jasenia-786.69217*lapsia+520.13951*desmod; (R-Square=0.3744)

Table 1 shows us how efficient this method is when compared with the original Toimtuki model. It tells that the TOIMTUKI has been improved to some extent, especially when we consider the absolute difference.

Table 1: Comparison between JUTTA and Linear Regression model

Model		JUTTA	Method 1
Variable		TUKI	TUKI
Number of Observation		1012	621
Absolute Error Percentage	[0,1)	0	0.00161
	[1, 10)	0.00296	0.01288
	[10, 100)	0.02569	0.05314
	[100, 1000)	0.33103	0.37037
	[1000, ∞)	0.64032	0.562
Relative Error Percentage	[0, 0.1%)	0	0.00322
	[0.1%, 1%)	0.00494	0.01771
	[1%, 10%)	0.03458	0.08535
	[10%, ∞)	0.96047	0.89372

Method two: 2-Stage Least Squares

Algorithm:

- Estimate the binary variable status, which describes the weather the person gets this benefit or not, meaning if he/she gets, then status=1, if he/she doesn't get, then status=0. This step is using Monte Carlo method. Firstly, by logistic regression, the estimated parameters are calculated, then by using $\pi_i = \exp(X\beta)/(1+\exp(X\beta))$, where π_i is the probability t of being status=1. Finally, generating random value from the uniform distribution, and compare this value with π_i the probability, if the random value is larger than the probability, giving status value 0, if not, giving value 1.
- Estimating the TUKI value by regression model in case the status=1, otherwise, give value 0. However, the estimated value could be negative, but in reality, it should be nonnegative value, so change the negative value to 0.

Table 2: Logistic estimated coefficients

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	-2.1593	0.1185	332.0326	<.0001
tyot	0.2023	0.0107	354.7804	<.0001
svatva	-0.00005	6.957E-6	59.0412	<.0001
maksvuok	0.00216	0.000184	137.2076	<.0001
vvvmk1	-0.00011	0.000019	35.2189	<.0001
desmod	-0.1751	0.0390	20.1498	<.0001
tyotseu	0.0429	0.0120	12.8695	0.0003
lpaktyva	0.00952	0.00287	10.9831	0.0009

So, by calculating $\pi_i = \exp(X\beta)/(1+\exp(X\beta))$, where β is the vector and its estimation has been shown in table 2. Next, a random number u_i is drawn from the standard uniform distribution, that is $u_i \sim U(0,1)$. Finally, by comparing u_i and π_i, we give the estimated status 1 and 0. In cases that the individual estimated status is 1, regression model is set to calculate the TUKI, while in other cases, TUKI will be given value 0 directly. Next step is the linear regression as showed in method one, by using equation (1), we could get:

$\hat{y}$=-152.15070+202.66061*tyot-0.07511*svatva+1.78485*maksvuok-0.13827*ttyotpr

+394.85153*jasenia+341.67733*desmod;(R-Square=0.3930)

Table 3 shows us how efficient this method is when compared with the original Toimtuki model.

Table 3: Comparison between JUTTA and 2SLS model

Model		TOIMTUKI	Method 2
Variable		TUKI	TUKI
Number of Observation		1012	894
Absolute Error Percentage	[0, 1)	0	0
	[1, 10)	0.00296	0.00224
	[10, 100)	0.02569	0.02685
	[100, 1000)	0.33103	0.42953
	[1000, ∞)	0.64032	0.54139
Relative Error Percentage	[0%, 0.1%)	0	0
	[0.1%, 1%)	0.00494	0.00224
	[1%, 10%)	0.03458	0.01454
	[10%, ∞)	0.96047	0.98322

From the table 3, we see that the second method is better than the original method in the absolute difference view, however, it is almost the same as the original method in the relative difference point of view.

Conclusion:

JUTTA model as a static microsimulation model performs quite well in all sub-models, only except for the “residual” model-Toimtuki. The Linear Regression model and 2SLS model both improved the accuracy of the Toimtuki to some extent, especially in the absolute difference percentage view, and there might be more potential significant variables which could help TOIMTUKI to be more accurate.

References:

TRIM3. The urban institute of US, 2012a. Available at: <*http://trim3.urban.org/T3IntroMicrosimulation.php*>. Accessed 2011.

TRIM3. The urban institute of US, 2012b. Available at: <*http://trim3.urban.org/documentation/Static%20versus%20Dynamic%20Microsimulation.html*>. Accessed 2011.

Anthony King, Hans Bækgaard, Martin Robinson(December 1999). *DYNAMOD-2: AN OVERVIEW*. Technical Paper no. 19, ISSN 1443-5098, ISBN 0858898004, in NATSEM, University of Canberra, Australia.

Lennart Flood, Fredrik Jansson, Thomas Pettersson, Tomas Pettersson, Olle Sundberg, Anna Westerberg(2005). *SESIM III- a Swedish dynamic micro simulation model*. In Handbook of SESIM051222, Swedish Ministry of Finance.

Rick Morrison(1998), Bernard Dussault(Edited 2000). *Overview of DYNACAN : a full-fledged Canadian actuarial stochastic model designed for the fiscal and policy analysis of social security schemes*. Statistics Cananda. Slightly adapted by Bernard Dussault(March 2000) for inclusion on the IAA website.

Honkanen Pertti(2010). *JUTTA-käsikirja*. Tulonsiirtojen ja verotuksen mikrosimulointijärjestelmä. Kela, Helsinki.

Appendix

htyotper:Basic unemployment allowance paid by KELA in Euros.

tyot: Number of month of person’s unemployment or forced leaving.

tyotseu: Number of month of person’s unemployment or forced leaving in year 2010.

ttyotpr: Unemployment allowance.

lpaktyva: Employee compulsory unemployment insurance.

vvvmk1: Paid earnings-related unemployment allowance.

vvvpvt1: Paid earning-related unemployment allowance days in total.

Workshop Programme

	Thursday 23 August	**Friday 24 August**	**Saturday 25 August**
		Chair: Mārtiņš Liberts	*Chair: Risto Lehtonen*
09:00 – 09:25		*Opening and Introduction:* *Mārtiņš Liberts*	KL: Monica Pratesi
09:25 – 09:50		IL: Gunnar Kulldorff	
09:50 – 10:15			KL: Monica Pratesi
10:15 – 10:40		C: Seppo Laaksonen D: Nicklas Pettersson	
10:40 – 11:05		*Refreshments*	*Refreshments*
		Chair: Daniel Thorburn	*Chair: Biruta Sloka*
11:05 – 11:30		C: Markus Gintas Šova D: Natalja Lepik	C: Daniel Thorburn D: Saara Oinonen
11:30 – 11:55		C: Ināra Kantāne D: Anna Larchenko	C: Jānis Kleperis D: Svetlana Badina
11:50 – 12:15		C: Oksana Honchar D: Katsiaryna Chystsenka	C: Ildikó Györki D: Riku Salonen
12:15 – 12:40		C: Olena Sugakova D: Meng Zhou	C: Julia Aru D: Mārtiņš Liberts
12:40 – 14:00		*Lunch*	*Lunch*
		Chair: Pauli Ollila	*Chair: Aleksandras Plikusas*
14:00 – 14:25		IL: Danute Krapavickaite	IL: Pauli Ollila
14:25 – 14:50			
14:50 – 15:15	*Registration of Participants* *(15:00 – 17:00)*	C: Aleksandras Plikusas D: Risto Lehtonen	C: Saara Oinonen D: Julia Orlova
15:15 – 15:40		C: Kaur Lumiste D: Markus Gintas Šova	C: Jelena Voronova D: Viktoras Chadyšas
15:40 – 16:05		*Refreshments*	*Refreshments*
		Chair: Natallia Bokun	*Chair: Ulrich Rendtel*
16:05 – 16:30		C: Viktoras Chadyšas D: Juris Breidaks	C: Katsiaryna Chystsenka D: Svitlana Slobodian
16:30 – 16:55		C: Julia Orlova D: Ināra Kantāne	C: Tomas Rudys D: Diana Santalova
16:55 – 17:20		C: Nicklas Pettersson D: Julia Aru	*Poster Session:* Juris Breidaks, Baiba Buceniece, Ance Ceriņa, Jānis Lapiņš, Tetiana Manzhos, Riku Salonen, Jeļena Vaļkovska
17:20 – 17:45		C: Maiken Mätik D: Olena Sugakova	
	Welcome Party (18:00)	*City Tour (18:00 – 20:00)*	*Steering Committee Meeting (19:00)*
KL – keynote lecture, IL – invited lecture, C – contributed paper, D – discussion			

	Sunday 26 August	**Monday 27 August**	**Tuesday 28 August**
	Chair: Olga Vasylyk	*Chair: Danute Krapavickaite*	*Chair: Jānis Lapiņš*
09:00 – 09:25	KL: Monica Pratesi	KL: Carl Erik Särndal	KL: Carl Erik Särndal
09:25 – 09:50			
09:50 – 10:15	C: Mauno Keto D: Kaur Lumiste	KL: Carl Erik Särndal	IL: Li-Chun Zhang
10:15 – 10:40	C: Vilma Nekrasaite-Liege D: Anastacia Bobrova		
10:40 – 11:05	*Refreshments*	*Refreshments*	*Refreshments*
		Chair: Li-Chun Zhang	*Chair: Imbi Traat*
11:05 – 11:30		IL: Anders Wallgren	C: Natalja Lepik D: Natallia Bokun
11:30 – 11:55	*Field Trip (11:30 - 20:00)*		C: Milda Šličkutė-Šeštokienė D: Lisha Wang
11:50 – 12:15		C: Inga Masiulaityte-Sukevic D: Oksana Honchar	C: Andris Fisenko D: Inga Masiulaityte-Sukevic
12:15 – 12:40		C: Anastacia Bobrova D: Tomas Rudys	C: Natallia Bandarenka D: Inta Priedola
12:40 – 14:00		*Lunch*	*Lunch*
		Chair: Seppo Laaksonen	*Chair: Anders Wallgren*
14:00 – 14:25		IL: Ulrich Rendtel	C: Mārtiņš Liberts D: Seppo Laaksonen
14:25 – 14:50			C: Lisha Wang D: Aleksandras Plikusas
14:50 – 15:15		C: Olga Vasylyk D: Natalja Budkina	C: Meng Zhou D: Ildikó Györki
15:15 – 15:40		C: Biruta Sloka D: Tetiana Manzhos	C: Natallia Bokun D: Andris Fisenko
15:40 – 16:05		*Refreshments*	*Refreshments*
		Chair: Markus Gintas Šova	*Chair: Gunnar Kulldorff*
16:05 – 16:30		C: Svitlana Slobodian D: Imbi Traat	C: Anna Larchenko D: Vilma Nekrasaite-Liege
16:30 – 16:55		*Round Table Discussions*	*Evaluation and Closing*
16:55 – 17:20			
17:20 – 17:45			
			Farewell Party (19:00)

KL – keynote lecture, IL – invited lecture, C – contributed paper, D – discussion

List of participants

First Name	Surname	Country	Institution	e-mail
Julia	Aru	Estonia	Tartu University	julia.aru@gmail.com
Svetlana	Badina	Norway	Statistics Norway	sbadyina@yahoo.com
Ilze	Balode	Latvia	Ventspils University College	ilze.balode@venta.lv
Natallia	Bandarenka	Belarus	Belarus State Economic University, Department of Statistics	bondnata@mail.ru
Anastacia	Bobrova	Belarus	Institute of Economic	nastassiabobrova@mail.ru
Natallia	Bokun	Belarus	Belarus State Economic University	nataliabokun@rambler.ru
Juris	Breidaks	Latvia	Central Statistical Bureau of Latvia	juris.breidaks@csb.gov.lv
Baiba	Buceniece	Latvia	Central Statistical Bureau of Latvia	baiba.buceniece@csb.gov.lv
Natalja	Budkina	Latvia	Riga Technical University	natalja.budkina@rtu.lv
Ance	Ceriņa	Latvia	Central Statistical Bureau of Latvia	ance.cerina@tvnet.lv
Viktoras	Chadyšas	Lithuania	Vilnius Gediminas Technical University	viktoras.chadysas@vgtu.lt
Katsiaryna	Chystsenka	Belarus	National Bank of Republic of Belarus	katsiaryna.chystsenka@gmail.com
Andris	Fisenko	Latvia	Central Statistical Bureau of Latvia	andris.fisenko@csb.gov.lv
Aleksandra	Galahina	Latvia	TNS Latvia	aleksandra.galahina@tns.lv
Ildikó	Györki	Hungary	Hungarien Central Statistical Office	ildiko.gyorki@ksh.hu
Oksana	Honchar	Ukraine	National Academy Statistics, Accounting and Audit	ohonchar@list.ru
Tetiana	Ianevych	Ukraine	Taras Shevchenko National University of Kyiv	yakovenkot@gmail.com
Ināra	Kantāne	Latvia	University of Latvia	inara.kantane@lu.lv
Mauno	Keto	Finland	Mikkeli University of Applied Sciences	mauno.keto@mamk.fi
Janis	Kleperis	Latvia	Institute of Solid State Physics of University of Latvia	kleperis@latnet.lv
Danutė	Krapavickaitė	Lithuania	Statistics Lithuania, Vilnius Gediminas Technical University	Danute.Krapavickaite@vgtu.lt
Gunnar	Kulldorff	Sweden	University of Umeå	gunnar@matstat.umu.se
Seppo	Laaksonen	Finland	University Helsinki	Seppo.Laaksonen@Helsinki.Fi
Janis	Lapins	Latvia	Bank of Latvia	Janis.Lapins@bank.lv
Anna	Larchenko	Belarus	Belarus State Economic University	human_clay-999@mail.ru
Risto	Lehtonen	Finland	University of Helsinki	risto.lehtonen@helsinki.fi
Natalja	Lepik	Estonia	Institute of Mathematical Statistics, University of Tartu	natalja.lepik@ut.ee
Mārtiņš	Liberts	Latvia	Central Statistical Bureau of Latvia, University of Latvia	martins.liberts@gmail.com
Kaur	Lumiste	Estonia	University of Tartu	kaur.lumiste@ut.ee
Tetiana	Manzhos	Ukraine	Kyiv National Economic University	tmanzhos@gmail.com

First Name	Surname	Country	Institution	e-mail
Valdis	Masalskis	Latvia	Bank of Latvia	Valdis.Masalskis@bank.lv
Inga	Masiulaityte-Sukevic	Lithuania	Statistics Lithuania	inga.masiulaityte@stat.gov.lt
Maiken	Mätik	Estonia	Tartu University	maiken.matik@gmail.com
Vilma	Nekrasaite-Liege	Lithuania	Vilnius Gediminas Technical University	nekrasaite.vilma@gmail.com
Saara	Oinonen	Finland	Statistics Finland	saara.oinonen@stat.fi
Pauli	Ollila	Finland	Statistics Finland	pauli.ollila@stat.fi
Julia	Orlova	Belarus	Belarus State Economic University	55xx@mail.ru
Nicklas	Pettersson	Sweden	Stockholm University	nicklas.pettersson@stat.su.se
Aleksandras	Plikusas	Lithuania	Vilnius University	Aleksandras.Plikusas@mii.vu.lt
Monica	Pratesi	Italy	University of Pisa	m.pratesi@ec.unipi.it
Inta	Priedola	Latvia	TNS Latvia	ipriedola@gmail.com
Ulrich	Rendtel	Germany	Freie Universitaet Berlin	Ulrich.Rendtel@fu-berlin.de
Tomas	Rudys	Lithuania	Statistics Lithuania	tomas.rudys@gmail.com
Riku	Salonen	Finland	Statistics Finland	riku.salonen@stat.fi
Diana	Santalova	Estonia	University of Tartu	Diana.Santalova@ut.ee
Carl-Erik	Sarndal	Sweden	Orebro University	carl.sarndal@telia.com
Maija	Skenderska	Latvia	Bank of Latvia	Maija.Skenderska@bank.lv
Svitlana	Slobodian	Ukraine	Vasyl Stefanyk Precarpathian National University	slobodian_s@ukr.net
Biruta	Sloka	Latvia	University of Latvia	Biruta.Sloka@lu.lv
Ieva	Strele	Latvia	Riga Stradins University	ievastrele@inbox.lv
Olena	Sugakova	Ukraine	Kyiv National University	sugak@univ.kiev.ua
Milda	Šličkutė-Šeštokienė	Lithuania	Statistics Lithuania	milda.slickute@stat.gov.lt
Markus Gintas	Šova	United Kingdom	Office for National Statistics	markus.sova@ons.gov.uk
Daniel	Thorburn	Sweden	Stockholm University	Daniel.thorburn@stat.su.se
Imbi	Traat	Estonia	University of Tartu	imbi.traat@ut.ee
Jeļena	Vaļkovska	Latvia	Central Statistical Bureau of Latvia	Jelena.Valkovska@csb.gov.lv
Olga	Vasylyk	Ukraine	Taras Shevchenko National University of Kyiv	olva75@gmail.com
Jelena	Voronova	Latvia	Central Statistical Bureau of Latvia	jelena.voronova@csb.gov.lv
Anders	Wallgren	Sweden	Örebro university	ba.statistik@telia.com
Britt	Wallgren	Sweden	Örebro university	ba.statistik@telia.com
Lisha	Wang	Sweden	Örebro University	lisha.wang@oru.se
Li-Chun	Zhang	Norway	Statistics Norway	lcz@ssb.no
Meng	Zhou	Finland	University of Helsinki	meng.zhou@helsinki.fi