DISSERTATIONES
MATHEMATICAE
UNIVERSITATIS
TARTUENSIS
123

KAUR LUMISTE

Improving accuracy of survey estimators by using auxiliary information in data collection and estimation stages

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

123

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
123

KAUR LUMISTE

Improving accuracy of survey estimators by using auxiliary information in data collection and estimation stages

UNIVERSITY OF TARTU
Press

Institute of Mathematics and Statistics, Faculty of Science and Technology, University of Tartu, Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in mathematical statistics on November 21, 2017, by the Council of the Institute of Mathematics and Statistics, Faculty of Science and Technology, University of Tartu.

Supervisor:

Associate Professor Imbi Traat, PhD
University of Tartu
Tartu, Estonia

Opponents:

Associate Professor Jan-Anton Grafström, PhD
Swedish University of Agricultural Sciences
Umeå, Sweden

Senior Statistician Peter Matthias Lundqvist, PhD
Statistics Sweden
Stockholm, Sweden

Commencement will take place on January 25, 2018 at 11:00 in J. Liivi 2-122 Tartu, Estonia.

The publication of this dissertation was financed by the Institute of Mathematics and Statistics, University of Tartu.

ISSN 1024-4212
ISBN 978-9949-77-644-3 (print)
ISBN 978-9949-77-645-0 (PDF)

University of Tartu Press
www.tyk.ee

To Marjeta, Erik, and Henrik,
for bringing joy into each and every day of my life, ...

... and Depeche Mode,
for making such bloody good music and
helping me pass many fruitful hours with the thesis.

Contents

Acknowledgements

My deepest and sincerest gratitude goes to my supervisor Associate Professor Imbi Traat. Her never ending optimism, her constant support and encouragement have helped to reach the finish of my doctoral studies. I am grateful for her countless suggestions and ideas in forming the dissertation.

Thank you, Professor Carl-Erik Särndal, for pointing towards today's problems facing survey sampling and responsive designs, and giving base material for the thesis. Thank you for the many lively discussions and sharing your expertise. It has always been a privilege to work with you.

I would like to thank my recently passed grandfather, Professor Ülo Lumiste. He was always routing for me in my studies, although our subjects were very far apart – differential geometry and survey sampling. Fortunately the news of the finished dissertation manuscript reached his hears before he passed away, I could see happiness from his weary eyes at that moment.

My gratitude also goes to all the professors, lecturers and colleagues at the Institute of Mathematics and Statistics. The knowledge and practices from the many lectures, seminars and private conversations have contributed in one way or the other to completion of the dissertation.

Finally, I would like to thank my family, for their patience, and my parents, for encouraging me to pursue an academic degree. Their support has helped me past many obstacles and difficult times.

List of original publications

The thesis is based on the following papers:

I Särndal, C.-E., Lumiste, K., and Traat, I. (2016). Reducing the Response Imbalance: Is the Accuracy of the Survey Estimates Improved? *Survey Methodology*, 42 (2): 219–238.

II Lumiste, K. (2017). Effect of auxiliary information in data collection and estimation stage. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 21 (1): 111–128.

III Lumiste, K. (2011). Consistent Estimation of Cross-classified Domains. *Statistics in Transition – new series*, 12 (2): 253–264.

Source of Paper I is Statistics Canada, Survey Methodology, November 2017. Reproduced and distributed on an "as is" basis with the permission of Statistics Canada. Paper II is reprinted with the kind permission from Acta et Commentationes Universitatis Tartuensis de Mathematica. Paper III is reprinted with the kind permission from Statistics in Transition new series.

Paper I is a joint paper with 2 coauthors. My major contribution was planning and executing an extensive simulation study. Description of this study and several illustrations of the results are summarised on more than 4 pages of the paper. I also participated and contributed in the theoretical discussions.

Other publications

- Yur'yev, A., Yur'yeva, L., Värnik, P., Lumiste, K., and Värnik, A. (2015). The Complex Impact of Risk and Protective Factors on Suicide Mortality: A Study of the Ukrainian General Population. *Archives of Suicide Research*, 19 (2): 249–259.
- Ainsaar, M., Lilleoja, L., Lumiste, K., and Roots, A. (2013). *ESS Mixed Mode Experiment Results in Estonia (CAWI and CAPI Mode Sequential Design)*, Tartu, University of Tartu.
- Yur'yev, A., Värnik, P., Sisask, M., Leppik, L., Lumiste, K., and Värnik, A. (2011). Some aspects of social exclusion: Do they influence suicide mortality? *International Journal of Social Psychiatry*, 59 (3): 232–238.

1. Introduction

1.1 Background

In sample surveys (or shortly surveys), the ultimate goal is to estimate unknown parameters of a population of interest, based on a selected sample from that population. Statistical results that the government and authorities of a country, researchers and other people daily consume are often coming from sample surveys. For example, news headlines announce the latest poll results on people's political party preference, politicians explain recent dynamics of nation's wellbeing, ministries determine new quotas for fishing, land use or forest management, companies make marketing decisions using customer satisfaction surveys.

The survey environment is rapidly changing today. New flexible designs are developed, that use emerging possibilities. Surveys are designed to use multiple data sources for sample selection, to collect data in multiple modes like through telephone calls and online web forms simultaneously, to achieve a better-balanced or representative response set by adaptive designs. Rich administrative data sources enable statistical agencies to publish register based statistics. There are also new innovative ways of approaching populations and obtaining data, like using mobile positioning and other big-data recorders. In the age of information, the speed of data production is a challenge, and the public always demands more and more accurate, timely, and relevant statistics. In this context, surveys are ever more challenging and complex to implement. Applying modern possibilities for improving survey designs is largely forced by increasing problems in the survey industry like declining response rates (i.e. high non-response rates), limited survey budgets and high response burdens of persons and companies, among others.

In a survey, the study variables (the variables whose unknown parameters like totals, means, proportions and other characteristics are required) are measured and respective data is used for estimation. There is yet another important type of variables – auxiliary variables. These variables play a crucial role in all steps of the survey process. They are used in most of the modern survey designs and data collection methods; their ultimate use is in the estimation stage. The overall aim of using auxiliary variables is to reduce bias and decrease variance of estimators. Fortunately the number of possible sources for auxiliary information is growing.

Data sources for auxiliary information. With advancements in technology and data storage becoming more affordable, many national statistical agencies and governments over the world have invested in digital registers and systems that provide data about many different populations within a country. For example, there are housing, business, population and motoring registers. The reasons behind creation of such databases are usually easier cost efficient governance and public openness, but these databases are also valuable for survey statisticians. For example, Finland maintains 3 key registers, like population, real estate and business registers, that are central to the functioning of the society, 4 other major registers like the taxation, employment, job applicants and pension registers, and other smaller and regional registers (Statistics Finland, 2004, pp. 10-12). Statistics Finland is allowed to use these rich sources of auxiliary information for statistics production.

Modern digital society has lead to wider adoption of computers and web forms in conducting surveys. This allows to gather paradata, i.e. administrative data about the data collection process. Kreuter (2013) gives a thorough overview of paradata. Paradata has proven useful for managing existing surveys and designing future ones (Couper, 1998; Chun and Kwanisai, 2010; Couper and Kreuter, 2012). For example, it can be number of contact attempts of each sampled element or data on interviewers who carry out fieldwork. European Social Survey interviewers, upon a visit gather data on survey respondents' housing and environment, irrelevant if the person fills in the survey or not. This data is used to evaluate measurement and non-response bias, and also used in the planning of next waves of data collection (Stoop et al., 2010).

After the completion of a survey project and publishing results the data is not thrown away. Many survey agencies store raw survey data to form their "pseudo" populations for quality control, lessening respondent burden and use recent survey results to validate or unify estimates of ongoing surveys. For example, Statistics Netherlands constructed a Social Statistical Database, where registers and data from sample surveys are linked. Estimates related to social statistics are obtained from this database (Houbiers, 2004).

All these mentioned sources provide an abundant pool of auxiliary information. Ideally this information can be in the form of known values for all population elements, but more commonly the information is known only for all sampled elements and/or in the form of known population totals. For example, national population registers often have data on age, gender, level of education, owning a driving license, and/or marital status among others, which can be used by survey agencies in different stages of the survey process.

Uses of auxiliary information. Auxiliary variables can be used in the sample selection stage to rule out "obscure samples". In the case of simple random sampling (SRS) of size n every sample of size n has a positive probability of occurring and that means a sample consisting of only males from a nation's population has a positive probability, although very small. Stratified and cluster sampling designs, probability proportional to size (usually noted πps) designs and balanced sampling designs use auxiliary variables to make samples more representative, where balanced means that auxiliary variable means in the sample and population are equal. Theory of stratified, cluster, πps and other sampling designs is thoroughly presented in Särndal et al. (1992). Methods which produce balanced samples given a set of auxiliary variables include the Cube Method (Deville and Tillé, 2004) and the Local Pivotal Method (Grafström et al., 2012). The use of auxiliary variables in the sampling process will not be discussed in current thesis, the focus is kept on the steps that follow sample selection – data collection and estimation stages.

In responsive and adaptive designs data collection strategies (i.e. treatments) are adapted to auxiliary information that becomes available before or during data collection. Monitoring the data collection process and planning interventions can bring a more appropriate final set of respondents, compared with a stationary design where the data collection follows a fixed unchanging protocol from beginning to end. Usually high response rates are demanded with any means, but Groves (2006) raised concerns about the quality of survey data with blindly gathering respondents. It was in response to these concerns that responsive and adaptive designs were born, with Groves and Heeringa (2006) being an early reference and a review on the literature is given by Tourangeau et al. (2017). There is practical evidence that responsive designs lower bias caused by non-response (e.g. Schouten et al., 2009; Schouten et al., 2012; Lundquist and Särndal, 2013; Särndal and Lundquist, 2014a; Särndal and Lundquist, 2014b; Särndal and Lundquist, 2017). In this thesis theoretical evidence is presented.

After data collection survey statisticians rely on estimation theory to resolve the challenge of non-response, primarily how to achieve low bias in the estimates. After the completion of data collection, the set of responding units is fixed. The choice of auxiliary variables plays a crucial role, an aspect that has been dealt with extensively, as in Särndal and Lundström (2005) and a comprehensive review of non-response weighting adjustment is found in Brick (2013). But assume that in the estimation stage additional auxiliary information is made available, for example, in the form of paradata or other register data. So the auxiliary vectors used in data collection and estimation stages may differ from each other. The aspect of different auxiliary vectors has been considered in Särndal and Lundquist (2014b), but receives more focused attention in current thesis.

After estimation stage the gathered information is still usable to improve quality of results in other surveys. Assume that population totals for certain variables are estimated in one survey, that we call the reference survey (RFS). Estimates of the same variables are produced in another survey, that we call the present survey (PRS), but in greater detail. In PRS the estimates are produced for totals of population subgroups, referred to as domains. The two surveys are carried out independently. Naturally, consumers would assume that these two surveys produce numerically consistent estimates - e.g. domain total estimates in PRS sum up to the population total estimate in RFS. Sadly this does not always hold. There are many methods to achieve consistency between estimates. Knottnerus (2003) proposes a general restriction estimator that is constructed upon unbiased initial estimators so that the result satisfies desired linear restrictions. Lepik (2011) extends this method for domain estimation. Särndal and Traat (2011) treat the inconsistency problem in another way and they propose a new method, the AC-calibration (A - auxiliary variables, C - common variables). Statistics Netherlands has also studied the problem and several articles have been published on repeated weighting (RW). Current thesis uses the last two methods to achieve consistency in a more complex case, i.e. for cross-classified domains under presence of outside sources (RFS) for marginal domains.

1.2 Aims of the Thesis

The general aim of this thesis is to contribute to the estimation theory in sample surveys by using auxiliary information. Auxiliary information is used for balanced data collection as well for calibration in the estimation phase. The broad question is whether this information, used in both phases, can increase accuracy of survey estimators more than its traditional application, i.e. using auxiliary information only in the estimation phase. The aim is to quantify the additional effect from balancing; also to characterise the contribution into final estimator if different auxiliary vectors are used in the two phases. Another question studied and solved in this thesis is consistency of survey estimates with other known results. Now auxiliary information is extended to involve also known information about study variable.

In more detail the goals are:

- Assuming there is considerable non-response, consider data collection methods that improve balance of the final response set with respect to selected auxiliary variables. The aim is to present theoretical evidence that balancing efforts during data collection will improve the accuracy, primarily reduce bias of the estimates that are ultimately produced by calibrated weighting (Paper I);
- Looking for the answer whether to use auxiliary information only for balancing response or only for calibrating estimators or for doing both, a novel situation with different auxiliary vectors in monitoring and estimation phases, is considered. Assume that after monitoring phase there is more auxiliary information available. The aim is to present the calibration estimator in a way that the effect of added auxiliary variables can be explicitly separated (Paper II);
- Assuming that some general information on study variable(s) is known from other surveys or registers, the aim is to develop formulas for consistent estimation of domains. The more complex situation of cross-classified domains is considered. The requirement is that the new domain estimators satisfy the summation restrictions, i.e. the estimators sum up to the known (or *a priori* estimated) marginal domain totals coming from external sources (Paper III);
- Illustrate and confirm thesis results in simulation studies.

1.3 Thesis outline

The thesis is arranged as follows: Section 2 introduces basics about design based survey sampling and notation, presents methods to employ auxiliary information, and gives formulas for estimators under full response and non-response. Section 3 includes short summaries of the contributions based on the three papers making up the thesis. Section 4 gives concluding remarks and discusses on further research ideas emerging from the considered topics.

2. Notation and methods

Throughout the thesis we consider the design-based approach where values of variables in a finite population are fixed constants and randomness is introduced by a *sampling design* – a set of rules and procedures by which elements of the finite population are included into the sample.

2.1 Design based sampling

Let $U = \{1, 2, \ldots, N\}$ denote a finite *population* of N units and the enumeration of population units is referred to as a *sampling frame*. Let a random vector (design vector) $\mathbf{I} = (I_1, I_2, \ldots, I_N)$ describe the sampling process on U and the random variable I_k indicates the number of selections for unit $k \in U$. For without-replacement (WOR) designs $I_k \in \{0, 1\}$ and with-replacement (WR) designs $I_k \in \{0, 1, 2, \ldots\}$. The multivariate distribution of the vector $\mathbf{I}$ with the probability function $\mathbf{I} \sim p(\mathbf{i}) = P(\mathbf{I} = \mathbf{i})$, where $\mathbf{i} = (i_1, i_2, \ldots, i_N)$ is a outcome of design vector $\mathbf{I}$, gives probabilistic description of the sampling design (Traat et al., 2004). In this thesis we consider WOR designs.

A sample s of size n is selected to estimate some characteristics of U according to a chosen sampling design. The sampling design generates for each element k a known inclusion probability, $P(k \in s) = E(I_k) = \pi_k > 0$, and a corresponding design weight $d_k = 1/\pi_k$. Designs where each unit in the population has exactly the same inclusion probability are called *self-weighting designs*. In case of non-response, data can only be collected from a subset r within the sample, $r \subset s \subset U$, and the values y_k of the study variable y are recorded for units $k \in r$ only.

Example 2.1. *An example of a simple random sampling design (SRS WOR). Let us have a population of* $N = 100$ *persons. Put 100 balls with persons' names into a bowl. Select 10 balls randomly from the bowl to form a sample of size* $n = 10$*. Inclusion probability for each person in this case would be* $\pi_k = n/N = 1/10$*.*

Example 2.2. *An example of a Bernoulli sampling design. For every person in a population of* $N = 100$ *a dice is rolled and if the result is a "6", then the person is included in the sample, otherwise not. Inclusion probability for each person in this case would be* $\pi_k = \pi = 1/6$*. This is a sampling design where the sample size is not fixed and on average the sample size would be* $E(n) = E\left[\sum_U I_k\right] = N\pi = 100/6$*.*

The objective is to estimate the population total $Y = \sum_U y_k$ of the study variable y, where y_k denotes value of the study variable for unit $k \in U$ (here and later $\sum_A$ denotes a sum over all the units k in set A).

2.2 Domains

Often insights are needed for population subgroups like geographical areas of households, social groups of people, economic field of operations of companies or soil type of farmlands. Let population U be divided into D non-overlapping and exhaustive domains U_d, $d \in \mathcal{D} = \{1, 2, \ldots, D\}$. Let domain indicators be

$$\gamma_k^d = \begin{cases} 1, & k \in U_d, \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

Domain totals of study variables can now be expressed as sums over the population

$$Y_d = \sum_{U_d} y_k = \sum_U \gamma_k^d y_k.$$

This reveals that the estimation results for population totals can be directly applied for the domain estimators.

Example 2.3. *Using the setting of Example 2.1, let the population register of* $N = 100$ *persons have the information of persons' gender, forming domains of males and females. Let each of the balls with a name be colored according to gender of the person it represents. Again, 10 balls are selected randomly WOR and information gathered from domain representatives in the sample is generalized to the whole domain in the estimation phase.*

Domains can be *a priori identified* (Example 2.3) or *unidentified.* Unidentified domains means that an element's classification into a domain is not known until it is determined from response. This case occurs if gender is not given in the population register.

2.3 Auxiliary information

It is assumed that there is access to auxiliary information on unit level, i.e. the vector of J auxiliary variables $\mathbf{x}_k = (\mathbf{x}_{1k,}\mathbf{x}_{2k}, \ldots, \mathbf{x}_{Jk})'$ is known at least for every element $k \in s$ (or for every $k \in U$ if it is compiled from comprehensive registers). Often, instead of population level individual values the known totals $\sum_U \mathbf{x}_k = \mathbf{X}$ belong to available auxiliary information. In this thesis we assume that the auxiliary variable vector satisfies

$$\boldsymbol{\mu}'\mathbf{x}_k = 1, \forall k \in U, \text{ for some vector } \boldsymbol{\mu} \text{ not depending on } k. \tag{2.2}$$

It is not a major restriction as most vectors $\mathbf{x}_k$ in practice are of this kind or can easily be altered to satisfy (2.2), as Example 2.4 will demonstrate.

An important type of auxiliary vector is a group vector. It identifies membership of every unit k in one of J mutually exclusive and exhaustive population groups, so that $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$, where the only 1 indicates the unique group (out of J possible) to which k belongs.

Example 2.4. *Assume that we have a numerical (continuous or discrete) auxiliary variable x_k, then by taking $\mathbf{x}_k = (1, x_k)'$ and $\boldsymbol{\mu} = (1, 0)'$ satisfies the requirement* (2.2). *When $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$ is a group vector, then $\boldsymbol{\mu} = (1, 1, \ldots, 1)'$ satisfies the requirement.*

Example 2.5. *A group vector can be constructed by crossing all categorical auxiliary variables. Let there be information on gender (male / female), employment status (employed / unemployed) and highest level of education (low / middle / high). Then the group vector $\mathbf{x}$ is a vector with dimension $J = 2 \times 2 \times 3 = 12$. With more variables and more variable levels the vector can become very long, for example Särndal and Lundquist (2014b) used a $\mathbf{x}$-vector of dimension 14 with 256 possible values for a study of the Swedish Living Conditions Survey.*

2.4 Balance of the response set

The concept of balance has been often used in statistical literature with reference to an equality of means of certain variables for two sets of units, where one is the subset of the other. For example balanced sampling aims to give a random sample so that the means of a set of auxiliary variables are equal (or approximately equal) within the sample and the population. Such sampling methods include the Cube Method (Deville and Tillé, 2004), and Local Pivotal Method (Grafström et al., 2012). Here we look at measuring balance of auxiliary variables in the response set with respect to the sample.

With the given auxiliary vector $\mathbf{x}$, means can be calculated for the response, $\bar{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k / \sum_r d_k$, and for the sample, $\bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k / \sum_s d_k$. If $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, then the response set is said to be perfectly balanced on the given $\mathbf{x}$-vector. In practice this is usually not the case and the J-dimensional mean difference $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ signals drift from perfect balance. A univariate indicator of imbalance is defined as

$$IMB = P^2 \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \Sigma_s^{-1} \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right), \tag{2.3}$$

where $\Sigma_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' / \sum_s d_k$ is a $J \times J$ weighting matrix assumed non-singular and $P = \sum_r d_k / \sum_s d_k$ is the design weighted response rate. IMB is a value between $0 \leq IMB \leq P(1-P)$ and can be calculated at any point during data collection. It characterizes r in relation to s with respect to the chosen $\mathbf{x}$-vector.

Based on the IMB value, or some other imbalance characteristic value, steps can be taken to change the initial data collection process with the aim to get a more balanced response in the end. For example, at fixed points in the data collection process, IMB is calculated from the gathered response. Hypothetical $IMB_{(k)}$ can be calculated for units $k \in s - r$, i.e. units who have not responded yet, if they were to be included into the response set. In the next data collection step we approach only those units that decrease the imbalance measure. The process is repeated in the next intervention point. The intervention points can, for example, depend on time or respondents in the response set (e.g. after collection of every 100 respondents). We call such activity monitoring the data collection process, and it is one example of a responsive design.

2.5 Estimation under full response

The basic design unbiased estimator of population total Y from a full sample is the Horvitz-Thompson (HT) estimator:

$$\hat{Y}_{FUL} = \sum_s d_k y_k. \tag{2.4}$$

For domain total $\hat{Y}_d$, it can be written $\hat{Y}_{dFUL} = \sum_s d_k \gamma_k^d y_k$, where γ_k^d is defined in (2.1) and $d \in \mathcal{D}$.

With the availability of auxiliary information, statistical agencies around the world use calibration estimators with the aim to get more accurate estimates. The principal behind calibration estimators is to adjust the design weights d_k with weights w_k so that the calibration equations $\sum_s d_k w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ would be satisfied. There are many ways for finding w_k, the two main methods are the distance minimization, used by Deville and Särndal (1992), and the instrument vector method considered in Estevao and Särndal (2000), and Kott (2006). The method with chi-square distance measure between weights d_k and w_k gives an analytic solution,

$$w_k = \left(\sum_U \mathbf{x}_k\right)' \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k.$$

These weights, using only auxiliary variables, are then applied for all study variables to get a calibration estimator,

$$\hat{Y}^*_{CAL} = \sum_s d_k w_k y_k. \tag{2.5}$$

Remark. *Notice that the weights w_k satisfy the population level calibration requirement:*

$$\sum_s d_k w_k \mathbf{x}_k' = \left(\sum_U \mathbf{x}_k\right)' \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_s d_k \mathbf{x}_k \mathbf{x}_k' = \sum_U \mathbf{x}_k'.$$

For domains the calibration estimator can be written

$$\hat{Y}^*_{dCAL} = \sum_s d_k w_k \gamma_k^d y_k,\ d \in \mathcal{D}.$$

2.6 Consistent estimation under full response

Assume that the population total Y is known, either from registers or accurately estimated from a previous survey (RFS), where the desired domains are unidentified and thus their estimation is impossible. In the present survey (PRS) the domain indicators are recorded and domain estimation becomes possible. It is natural to demand that the estimated domain totals in the PRS sum up to the corresponding known totals.

Example 2.6. *In the Example 2.3 with the population forming domains of males and females from register info, there might be a need to further investigate happy and unhappy males and females. This additional separation is not known until sampled persons are surveyed in PRS. So a survey is done regarding these new domains, but the domain estimates will probably be inconsistent with marginal information from RFS.*

Särndal and Traat (2011) proposed to use what they call AC-calibration (A - auxiliary and C - common variables, further the terms A-variables and C-variables are used). This method is basically standard calibration where the auxiliary information vector $\mathbf{x}_k$ is extended with C-variables. For each element $k \in s$, construct a $J+1$ dimensional vector:

$$\begin{pmatrix} \mathbf{x}_k \\ y_k \end{pmatrix}, k \in s.$$

The calibration weights are then

$$w_{ACk} = \begin{pmatrix} \mathbf{X} \\ Y_0 \end{pmatrix}' \mathbf{M}^{-1} \begin{pmatrix} \mathbf{x}_k \\ y_k \end{pmatrix}, \tag{2.6}$$

where $\mathbf{X} = \sum_U \mathbf{x}_k$, $Y_0 = \sum_U y_k$ or is its estimate $Y_0 = \hat{Y}$ from RFS, and

$$\mathbf{M} = \sum_s d_k \begin{pmatrix} \mathbf{x}_k \\ y_k \end{pmatrix} \begin{pmatrix} \mathbf{x}_k \\ y_k \end{pmatrix}'.$$

Plugging the AC-weights (2.6) into the calibration estimator we get

$$\hat{Y}_{CAL}^{AC} = \sum_s d_k w_{ACk} y_k,$$

and we can use the domain indicators (2.1) to find AC-calibration estimators for domains.

Kroese and Renssen (1999), Houbiers (2004), Knottnerus and Van Duin (2006) use the repeated weighting method (RW) to achieve consistency. The overall idea is to calibrate the initial calibration weights w_k again to get new weights w_{RWk}, so that $\sum_s d_k w_{RWk} y_k = Y_0$. The auxiliary vector in this case would be

$$\begin{pmatrix} 1 \\ y_k \end{pmatrix}, k \in s,$$

so that requirement (2.2) would be satisfied and RW-weights have the form:

$$w_{RWk} = \begin{pmatrix} \hat{N} \\ Y_0 \end{pmatrix}' \left(\sum_s d_k \begin{pmatrix} 1 \\ y_k \end{pmatrix} \begin{pmatrix} 1 \\ y_k \end{pmatrix}' \right)^{-1} \begin{pmatrix} 1 \\ y_k \end{pmatrix},$$

where $\hat{N} = \sum_s d_k$ estimates the population size.

Remark. *Notice that the new weights w_{RWk} will not satisfy the calibration requirements for $\mathbf{x}$-variables, i.e*

$$\sum_s d_k w_{RWk} \mathbf{x}_k \neq \sum_U \mathbf{x}_k.$$

2.7 Estimation under non-response

In the presence of non-response HT-estimators cannot be computed. In such a situation one can use the simple expansion estimator

$$\hat{Y}_{EXP} = \hat{N} \bar{y}_r, \tag{2.7}$$

where $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$ is the design weighted mean of the study variable in r. But this estimator is often considerably biased, so a more widely used method is to calibrate on the auxiliary vector $\mathbf{x}_k$:

$$\hat{Y}_{CAL} = \sum_r d_k g_k y_k,$$

where

$$g_k = \left(\sum_s d_k \mathbf{x}_k \right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_k \tag{2.8}$$

are the calibration weights under non-response (g-weights for short).
Remark. *Notice that the weights g_k satisfy the sample level calibration requirement*

$$\sum_r d_k g_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k,$$

where $\sum_s d_k \mathbf{x}_k$ are unbiased estimates for population totals $\sum_U \mathbf{x}_k$.

2.8 Imbalance of the study variable

We analyze the relationship between study variable imbalance, caused by non-response, and imbalance of the auxiliary variables. We measure y-variable imbalance with $\bar{y}_r - \bar{y}_s$, where $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$ is the design weighted mean of y in the response set, $\bar{y}_s = \sum_s d_k y_k / \sum_s d_k$ is the design weighted mean of y in the sample, and $\mathbf{x}$-vector imbalance with IMB defined in (2.3). By multiplying the difference $\bar{y}_r - \bar{y}_s$ with $\hat{N} = \sum_s d_k$, we are able to extract two meaningful terms:

$$\hat{N}(\bar{y}_r - \bar{y}_s) = \hat{Y}_{EXP} - \hat{Y}_{FUL} = (\hat{Y}_{EXP} - \hat{Y}_{CAL}) + (\hat{Y}_{CAL} - \hat{Y}_{FUL}). \quad (2.9)$$

The first term $\hat{Y}_{EXP} - \hat{Y}_{CAL}$ is computable and shows adjustment from the naive expansion estimator when we calibrate with auxiliary information. The second term $\hat{Y}_{CAL} - \hat{Y}_{FUL}$ is not computable and quantifies the deviation of calibration estimator under non-response from the full sample HT-estimator. This illustrates the deviations we face in estimation stage due to non-response.

The property $\boldsymbol{\mu}'\mathbf{x}_k = 1$ in (2.2) reveals that $\bar{\mathbf{x}}_r'\mathbf{b}_r = \bar{y}_r$, $\bar{\mathbf{x}}_s'\mathbf{b}_s = \bar{y}_s$ and $\bar{\mathbf{x}}_s'\mathbf{b}_r = \hat{Y}_{CAL}/\hat{N}$, where $\mathbf{b}_r$ and $\mathbf{b}_s$ are ordinary linear regression coefficient vectors for the whole sample s and for the response r, respectively,

$$\begin{aligned} \mathbf{b}_s &= \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_s d_k \mathbf{x}_k y_k\right), \\ \mathbf{b}_r &= \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_r d_k \mathbf{x}_k y_k\right). \end{aligned}$$

Now we get from (2.9)

$$\bar{y}_r - \bar{y}_s = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\,\mathbf{b}_r + (\mathbf{b}_r - \mathbf{b}_s)'\,\bar{\mathbf{x}}_s. \quad (2.10)$$

The decomposition (2.10) highlights two undesirable differences, $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ due to imbalance and $\mathbf{b}_r - \mathbf{b}_s$ due to inconsistent regression. The adjustment term $(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\,\mathbf{b}_r = (\hat{Y}_{EXP} - \hat{Y}_{CAL})/\hat{N}$ can clearly be reduced by constructing r to have low imbalance. The effects of low IMB on $(\mathbf{b}_r - \mathbf{b}_s)'\,\bar{\mathbf{x}}_s = (\hat{Y}_{CAL} - \hat{Y}_{FUL})/\hat{N}$ is not so evident and is studied further in the next section and Paper I.

3. Summary of contributions

3.1 Summary of Paper I

The paper explores the deviance of the calibration estimator $\hat{Y}_{CAL}$ from the unbiased (but unrealized) full sample estimator $\hat{Y}_{FUL}$. Under specified assumptions, the conditional expectation and variance of that deviance are derived.

In Result 1 of the paper, the design-based approach is chosen with y_k being non-random. A self-weighting sampling design of size n and equally probable response sets of size m are assumed. The auxiliary vector is assumed to be a group vector. In Result 2 the model-based approach is considered where random y_k are regressed on auxiliary vector $\mathbf{x}_k$ with assumptions on the residuals. No restrictions are made on the auxiliary variables.

The observed deviance is

$$\mathbf{\Delta}_r = (\hat{Y}_{CAL} - \hat{Y}_{FUL})/\hat{N} = (\mathbf{b}_r - \mathbf{b}_s)' \, \bar{\mathbf{x}}_s.$$

Under both approaches, the conditional expectation of $\mathbf{\Delta}_r$, for given r, s, and $\mathbf{x}$-variables, is zero. The exact formula for conditional variance of $\mathbf{\Delta}_r$ is derived. In Result 1 the approximation of that variance, explicitly showing the relationship with IMB, is given. Under assumptions where the sample s is divided into non-overlapping subgroups s_j of size n_j, $j \in \{1, 2, \ldots, J\}$, by the group vector $\mathbf{x}_k$, and where group variance S^2_{yj} and the group response rates $p_j = m_j/n_j$ vary by little only over the groups, the approximation is

$$S^2_{\Delta} = \mathrm{Var}\left(\mathbf{\Delta}_r | \bar{\mathbf{x}}_r, m, s\right) \approx \left(1 - p + \frac{IMB}{p^2}\right) \frac{S^2_y}{m}, \tag{3.11}$$

with $p = m/n$ as the overall response rate, and

$$\begin{aligned} S_y^2 &= \sum_{j=1}^{J} \frac{n_j}{n} S_{yj}^2, \\ S_{yj}^2 &= \sum_{s_j} \frac{\left(y_k - \bar{y}_{s_j}\right)^2}{n_j - 1}. \end{aligned}$$

Here we specify the approximation (3.11) by giving its remainder term. The formula (3.11) follows from the exact formula (7.3) in Paper I which is here presented in the form

$$S_\Delta^2 = \operatorname{Var}\left(\boldsymbol{\Delta}_r | \bar{\mathbf{x}}_r, m, s\right) = \frac{1}{m}\left[(1-p)S_y^2 + \sum_{j=1}^{J} \frac{n_j}{n}\left(\frac{p}{p_j} - 1\right) S_{yj}^2\right]. \quad (3.12)$$

Let us assume equal variances in groups, $S_{yj}^2 = S_y^2$, and study the term of (3.12),

$$\sum_{j=1}^{J} \frac{n_j}{n}\left(\frac{p}{p_j} - 1\right). \quad (3.13)$$

Denote

$$\delta_j = \left(\frac{p_j}{p} - 1\right), \quad (3.14)$$

then IMB, which we want to see in (3.13), has in our group vector case the form,

$$IMB = p^2 \sum_{j=1}^{J} \frac{n_j}{n}\left(\frac{p_j}{p} - 1\right)^2 = p^2 \sum_{j=1}^{J} \frac{n_j}{n} \delta_j^2.$$

Assuming $|\delta_j| < 1$, the power series expansion (Råde and Westergren, 1988, p.192) gives,

$$\frac{1}{1+\delta_j} = 1 - \delta_j + \delta_j^2 + R_j,$$

where

$$R_j = \frac{1}{1+\delta_j} - 1 + \delta_j - \delta_j^2 = -\frac{\delta_j^3}{1+\delta_j}.$$

Since

$$\frac{p}{p_j} - 1 = \frac{1}{1+\delta_j} - 1 = -\delta_j + \delta_j^2 + R_j,$$

we get for (3.13)

$$\sum_{j=1}^{J} \frac{n_j}{n}\left(\frac{p}{p_j} - 1\right) = \sum_{j=1}^{J} \frac{n_j}{n}(-\delta_j) + \sum_{j=1}^{J} \frac{n_j}{n}\delta_j^2 + \sum_{j=1}^{J} \frac{n_j}{n} R_j. \tag{3.15}$$

The first term on the right hand side of (3.15) is 0, the second term is IMB/p^2, and the third term is related to IMB, as we see below.

Now, under assumptions $S_{yj}^2 = S_y^2$ and $|\delta_j| < 1$ for each j, our formula (3.12) takes the form

$$S_{\Delta}^2 = \frac{S_y^2}{m}\left(1 - p + \frac{IMB}{p^2} + \sum_{j=1}^{J} \frac{n_j}{n} R_j\right). \tag{3.16}$$

Let us evaluate the absolute value of the remainder term:

$$\left|\sum_{j=1}^{J} \frac{n_j}{n} R_j\right| = \sum_{j=1}^{J} \frac{n_j}{n}\delta_j^2 \left|-\frac{\delta_j}{1+\delta_j}\right| \leq \frac{IMB}{p^2} \max_j \left|\frac{\delta_j}{1+\delta_j}\right|.$$

Comparing the last two terms in (3.16), we have

$$\left|\frac{\sum_{j=1}^{J} \frac{n_j}{n} R_j}{IMB/p^2}\right| \leq \max_j \left|\frac{\delta_j}{1+\delta_j}\right|. \tag{3.17}$$

Now, in the process $p_j \to p$, $\forall j$, we have $\delta_j \to 0$, $\forall j$, due to (3.14), and the bound in (3.17) goes to 0. Consequently, in this process, $\sum_{j=1}^{J} \frac{n_j}{n} R_j = o(\frac{IMB}{p^2})$, i.e. the remainder term in (3.16) goes to 0 faster than IMB/p^2. This justifies our approximation (3.11). Moreover, we have the explicit expression for the remainder term $\sum_{j=1}^{J} \frac{n_j}{n} R_j$, and can analyze its size for any fixed set of p_j-values.

The result (3.16) says that under conditions for p_j, by decreasing imbalance IMB of the response set, in general, we decrease conditional variance of $\boldsymbol{\Delta}_r$. Consequently there is smaller risk to have large deviation between calibration estimator and the unbiased full sample HT-estimator. We see that the right hand side of (3.16) is zero for full response when $p = 1$. Then $p_j = 1$ implies $\delta_j = 0$, $\forall j$, and IMB as well as R_j are zero. The factor S_y^2 decreases magnitude of $S_{\boldsymbol{\Delta}}^2$ when S_{yj}^2 are small, which happens under strong relationship between y- and $\mathbf{x}$-variables.

Result 2 in the paper makes a small exception from following the design based approach in the thesis. The result is derived under model-based approach, where the finite population values are assumed to be random with assumptions to their random nature. In Result 2 we assume a linear regression model between y and $\mathbf{x}$. The y_k are considered random, with properties stated by the model. A group vector feature for $\mathbf{x}$ is no longer necessary. The conclusions are in some respects similar to those in Result 1.

The theoretical results were validated in a simulation study with real data from an Estonian household survey. Illustration of the relationship (3.11) is given in Figure 3.1.

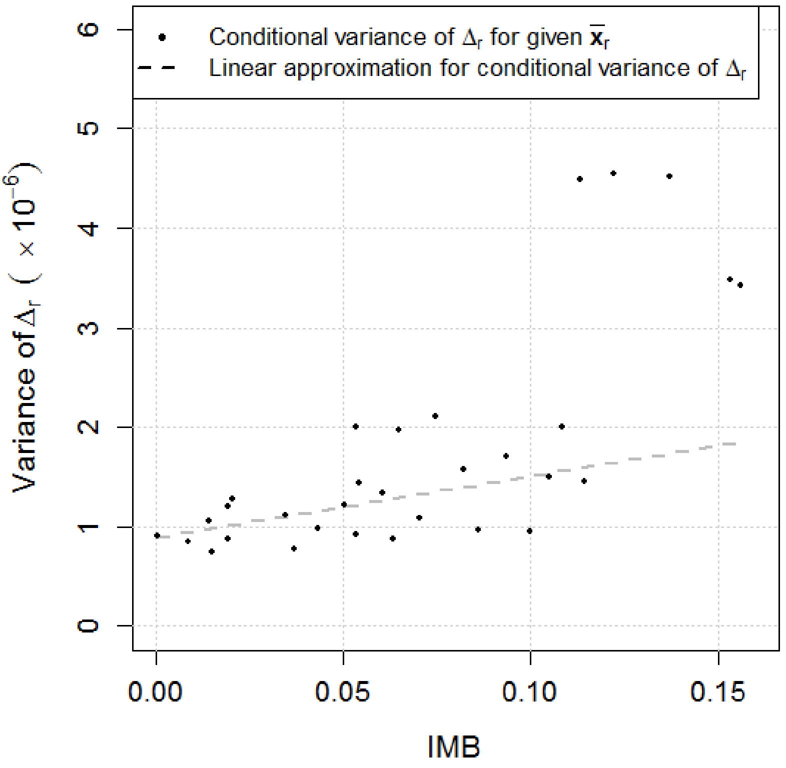

Figure 3.1: Conditional variance of $\mathbf{\Delta}_r$ as a function of imbalance *IMB*; $\mathbf{x}_k$ is a group vector of dimension 3; response sets r of fixed size 12 from a fixed sample s of size 20.

3.2 Summary of Paper II

In responsive survey designs the goal is to get a well representative set of respondents through planning and appropriate intervention in the data collection process. One possibility is to monitor data collection with more advanced quality measures than the response rate, here we consider the imbalance measure IMB, defined in (2.3), with respect to given auxiliary variables. Thus, monitoring response means computing IMB and pursuing respondents who would reduce imbalance in several time points during data collection. Literature found on IMB uses the same $\mathbf{x}$-vector in the monitoring and estimation stages, i.e. in calculation of IMB and definition of estimators that use auxiliary information.

Here a clear distinction is made, assume that we have access to additional calibration variables after data collection, for example in the form of paradata. Paradata, collected during data collection phase for each element in the sample, can be used for calibrating estimates from the response level to the sample level. Like in Paper I the decomposition (2.10) is taken under study, but the auxiliary vector is split into two parts:

$$\mathbf{x}_k = \mathbf{x}_{MEk} = \begin{pmatrix} \mathbf{x}_{Mk} \\ \mathbf{x}_{Ek} \end{pmatrix}, \qquad (3.18)$$

where $\mathbf{x}_{Mk} : p \times 1$ is an auxiliary vector used for monitoring response and $\mathbf{x}_{Ek} : q \times 1$ is an auxiliary vector of extra set of variables that are later added to compute the calibration estimator in the estimation stage. Here indexes M and E show whether the auxiliary vector $\mathbf{x}_{Mk}$ or $\mathbf{x}_{Ek}$ is used.

Using the new auxiliary vector (3.18) the g-weights in (2.8) can be split into two distinct parts:

$$g_k = g_{Mk} + h_k,$$

where g_{Mk} is dependent only on auxiliary information available or used for monitoring response and h_k is residual information from added auxiliary variables not linearly explained by $\mathbf{x}_M$-variables. The explicit expressions for these terms are derived in the paper. Useful properties of the new weights h_k are proved. The split weights g_k are used to express calibration estimator in two terms, $\hat{Y}_{CAL} = \sum_r d_k g_{Mk} y_k + \sum_r d_k h_k y_k$, the first one calibrating on monitoring variables. If full balance with respect to the $\mathbf{x}_M$-variables is received in the response set, then the first term is a simple expansion estimator $\hat{Y}_{EXP}$ in (2.7). The second term explicitly includes lack of balance indicator of added auxiliary variables, $\bar{\mathbf{x}}_{Er} - \bar{\mathbf{x}}_{Es}$, and the effect of residuals $\boldsymbol{\varepsilon}_k$ when regressing $\mathbf{x}_{Ek}$ on $\mathbf{x}_{Mk}$.

Auxiliary vector (3.18) also enables us to split IMB into two terms:

$$IMB = IMB_M + IMB_E,$$

where IMB_M is the imbalance measure with regard to $\mathbf{x}_M$-information and IMB_E is extra imbalance introduced by $\mathbf{x}_E$-information. IMB_E has a more complex formula behind it than the IMB in (2.3), see Paper II for the exact form.

A simulation study was carried to test two interesting cases of survey processes that emerged from theoretical part:

Case 1 Response accumulation is not monitored, instead additional auxiliary information is added in the estimation stage, i.e. $\mathbf{x}_M$-information and $\mathbf{x}_E$-information are only used to calculate the calibration estimator.

Case 2 Response accumulation is monitored and the process intervened to gather a response r with lower IMB value, but no extra auxiliary variables were used in estimation, i.e. only $\mathbf{x}_M$-information is used to calculate the calibration estimator.

That produced two response sets with response rate for both cases being 60%, calibrated estimates were found using only $\mathbf{x}_M$-variables in Case 2 and $\mathbf{x}_{ME}$-variables in Case 1. This was repeated 1000 times.

The results suggest that if $\mathbf{x}_E$-variables are not strongly correlated with the study variable, then monitoring and balancing the response set (Case 2) gives slightly more accurate results in terms of bias from the true population totals. If $\mathbf{x}_E$-variables are strongly correlated with the study variable, then we would get more accurate results in Case 1, compared to the estimation with monitored responses and no extra auxiliary information.

Balancing on the monitoring variables reduced IMB_E, implying that the lack of balance of added auxiliary variables, $\bar{\mathbf{x}}_{Er} - \bar{\mathbf{x}}_{Es}$, decreased.

3.3 Summary of Paper III

The novel results of Paper III concern the situation where we have two sources of information on the study variables (either surveys or registers). One source, let it be RFS1, has information on domains formed by certain categorical variable, not considered or not identified in the other source

RFS2. Instead, RFS2 has information on domains formed by another categorical variable. In the present survey PRS we are however interested in study variable estimates in domains cross-classified with these categorical variables. We are aiming to estimates that are consistent with known information.

In the Example 2.6 this would correspond to a situation where RFS1 has domains by males and females. Domains of happy and unhappy people is studied in RFS2, but domains are not separated by gender. In PRS we are interested in cross-classified domains of happy and unhappy males and females, and y total estimates that would be consistent with RFS1 and RFS2.

To achieve consistency AC-calibration and repeated weighting (RW) were used and formulas were developed for the previously explained case. In these formulas we have cross-classified domains in the form of $a \times b$ table, where a and b are the number of category levels of the variables forming domains in RFS1 and RFS2 respectively. The formulas are derived for the case of multiple study variables, i.e. $\mathbf{y}$ is a v-dimensional vector and we estimate $\mathbf{y}$-totals in table cells.

The AC-calibrated estimates that are consistent with known table marginals (sum up to these marginals), are given by the formula

$$\hat{\mathbf{Y}}^{AC}_{CAL} = \sum_{s} w_{ACk} \mathbb{I}_k \otimes \mathbf{y}_k,$$

where w_{ACk} are the calibrated weights on A- and C-information and they encompass the design weight d_k, $\mathbb{I}_k$ is an indicator matrix for the cross-classified domains. The result $\hat{\mathbf{Y}}^{AC}_{CAL}$ is a $(av) \times b$ matrix of domain total estimators for study variables $\mathbf{y}$ that satisfy marginal sums coming from external sources. The explicit form of the weights w_{ACk}, as well formal definitions and formulas of RW in the cross-classified case are given in Paper III.

The formulas of AC-calibration and RW for the cross-classified domains were tested in a simulation study. Simulations were done on a population composed of real data from the Estonian Household Survey. The results revealed that consistency was achieved with the information from RFS using AC-calibration and RW. In terms of comparing AC versus RW they both give almost equal estimates, compared to HT-estimates AC and RW estimates had lower variance, but slightly higher bias, since calibration methods are approximately unbiased.

4. Concluding remarks and open problems

As a conclusion, the aims of the thesis were achieved. Explicit expressions were derived highlighting the stated research problems.

A connection between imbalance and the deviance of the calibration estimator $\hat{Y}_{CAL}$ from the unbiased full information estimator $\hat{Y}_{FUL}$ was shown for two special cases.

In the case of different auxiliary variable vectors being used in the data collection and estimation stages, the expression for the calibration weights with two terms were derived, explicitly showing the contribution of $\mathbf{x}_M$- and $\mathbf{x}_E$-variables to the calibration estimator $\hat{Y}_{CAL}$. Useful properties were proved and special cases were studied where full balance with respect to $\mathbf{x}_M$-variables was assumed.

Formulas for achieving consistency in a case of cross-classified domains and known marginal information from two external sources, were derived. Simulation studies confirmed the results, and enabled further insight on previously mentioned topics.

Today there are many sources for auxiliary information and methods are being developed to use them in different stages of the survey process. The aims have been: (i) achieving more accurate estimates by reducing variance or reducing bias caused by non-response, or (ii) make the survey processes more efficient.

The question, whether to make balancing efforts during data collection phase, or only later, in the estimation stage, is still under research focus (Brick and Tourangeau, 2017).

In general, the author of this thesis advocates the use of responsive designs where data collection is monitored to arrive at a more representative response set. Many practical papers with same conclusions have been published, but there are few with theoretical evidence. So the search continues for further evidence of balancing effect on the non-response bias of calibration estimator, e.g. by relaxing assumptions of the results in Paper I.

An idea in similar direction would be to investigate study variable imbalance (2.10) when the full sample HT-estimator (2.4) is substituted with full sample calibration estimator (2.5) with population level information. It is widely known that the calibration estimator has lower variance than the HT-estimator, so with non-response we should try to aim for the full information calibration estimator. With this we loose the appealing property of reducing deviance form an unbiased estimator, but calibration estimators are approximately unbiased and statistical agencies around the world use calibration estimators instead of the HT-estimator.

Another open problem is present with consistent estimation with information from outside sources. Särndal and Traat (2011) and in the thesis we consider only the case where there is full response. A more realistic case to study is when non-response occurs and we still want to achieve consistency with known study variable totals. Also the case of cross-classified domains would be of interest.

References

Brick, J.M. (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of Official Statistics*, 29 (3): 329–353.

Brick, J.M. and Tourangeau, R. (2017). Responsive Survey Designs for Reducing Nonresponse Bias. *Journal of Official Statistics*, 33 (3): 735-752

Chun, Y. and Kwanisai, M. (2010). A Response Propensity Modeling Navigator for Paradata. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 356-369. Joint Statistical Meetings, Vancouver, Canada.

Couper, M.P. (1998). Measuring Survey Quality in a CASIC Environment. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, Washington, DC.

Couper, M.P. and Kreuter, F. (2012). Using Paradata to Explore Item Level Response Times in Surveys. *Journal of the Royal Statistical Society: Series A*, 176: 271-286.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimator in Survey Sampling. *Journal of the American Statistical Association*, 87: 376–382.

Deville, J.-C. and Tillé, Y. (2004). Efficient Balanced Sampling: The Cube Method. *Biometrika*, 91 (4): 893–912.

Estevao, V.M. and Särndal, C.-E. (2000). A Functional Approach to Calibration. *Journal of Official Statistics*, 16 (4): 379–399.

Grafström, A., Lundström, N. L., and Schelin, L. (2012). Spatially Balanced Sampling Through the Pivotal Method. *Biometrics*, 68: 514-520.

Groves, R. (2006). Research Synthesis: Nonresponse Rates and Nonresponse Error in Household Surveys. *Public Opinion Quarterly*, 70: 646–675.

Groves, R.M. and Heeringa, S.G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. *Journal of the Royal Statistical Society, Series A*, 169: 439–457.

Houbiers, M. (2004). Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. *Journal of Official Statistics*, 20 (1): 55-75

Knottnerus, P. (2003). *Sample Survey Theory. Some Pythagorean Perspectives.* Springer, New York.

Knottnerus, P. and van Duin, C. (2006). Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22: 565–584.

Kott, P.S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. *Survey Methodology*, 32 (2): 133–142.

Kreuter, F. (2013). *Improving Surveys with Paradata. Analytic Use of Process Information.* John Wiley & Sons, Inc., Hoboken.

Kroese, A.H. and Renssen, R.H. (1999). Weighting and Imputation at Statistics Netherlands. In *Proceedings of IASS Satellite Conference on Small Area Estimation*, Riga, Latvia: 109–120.

Lepik, N. (2011). *Estimation of Domains Under Restrictions Built Upon Generalized Regression and Synthetic Estimators.* Dissertation, University of Tartu.

Lundquist, P. and Särndal, C.-E. (2013). Aspects of Responsive Design With Applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29: 557–582.

Råde, L. and Westergren, B. (1988). *Mathematics Handbook for Science and Engineering.* Studentlitteratur, Lund.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., and Skinner, C. (2012). Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators. *International Statistical Review*, 80 (3): 382–399.

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the Representativeness of Survey Response. *Survey Methodology*, 35 (1): 101–113.

Statistics Finland (2004). *Use of Register and Administrative Data Sources for Statistical Purposes: Best Practices of Statistics Finland.* Ed. P. Myrskylä. Valopaino, Helsinki.

Stoop, I., Matsuo, H., Koch, A., and Billiet, J., (2010). Paradata in the European Social Survey: Studying Nonresponse and Adjusting for Bias. In *Proceedings of the 2010 Joint Statistical Meetings, Section on Survey Research Methods*, Vancouver, Canada: 407–421.

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

Särndal, C.-E. and Lundquist, P. (2014). Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation. *Journal of Survey Statistics and Methodology*, 2: 361–387.

Särndal, C.-E. and Lundquist, P. (2014). Balancing the Response and Adjusting Estimates for Nonresponse Bias: Complementary Activities. *Journal de la Société Française de Statistique*, 155 (4): 28–50.

Särndal, C.-E. and Lundquist, P. (2017). Inconsistent Regression and Nonresponse Bias: Exploring Their Relationship as a Function of Response Imbalance. *Journal of Official Statistics*, 33 (3): 709–734.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* John Wiley & Sons, Inc., New York.

Särndal, C.-E. and Traat, I. (2011). Domain Estimators Calibrated on Information from Another Survey. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 15 (2): 43–60.

Tourangeau, R., Brick, J.M., Lohr, S., and Li, J. (2017). Adaptive and Responsive Survey Designs: a Review and Assessment. *Journal of the Royal Statistical Society, Series A*, 180 (1): 203–223.

Traat, I., Bondesson, L., and Meister, K. (2004). Sampling Design and Sample Selection Through Distribution Theory. *Journal of Statistical Planning and Inference*, 123 (2): 395–413.

PAPERS

Kokkuvõte

Valikuuringute hinnangute täpsuse tõstmine abiinformatsiooni kasutamisega andmekogumise ja hindamise etappides

Valikuuringute eesmärgiks on hinnata üldkogumi parameetreid juhusliku valimi pealt. Igapäevaselt kuuleme uudiseid ja otsuseid, mille aluseks on valikuuringute tulemused. Näiteks võime ajalehtede pealkirjadest lugeda rahva poliitiliste parteide eelistustest või ministeeriumi ametite kehtestatud kalapüügi ja metsaraie kvootidest.

Valikuuringute keskkond on aga pidevas muutuses ja arenev. Tänu tehnika arengule on võimalik läbi viia keerukaid valikuuringuid, mis kasutavad mitmeid andmeallikaid tulemuste ja protsesside parandamiseks, sealhulgas uudseid mobiilpositsioneerimise, sensorandmete ja muude suurandmete ("*Big Data*") salvestisi. Administratiivandmed võimaldavad statistika ametitel avaldada registrite põhiseid näitajaid.

Uudsete meetodite rakendamise ja arendamise vajadus on enamasti tingitud valikuuringute keskkonda vaevavate probleemide tõttu. Juba pikemat aega on valikuuringute vastamismäärad pidevas languses, küsitluste populaarsuse kasvu tõttu räägitakse juba vastajate kurnamisest. Uuringute läbiviimine on läinud üha kallimaks, samas kui eelarved on jäänud samaks või hoopis vähenenud. Mitmete uudsete meetodite keskmes on abiinformatsioon.

Valikuuringute eesmärgiks on hinnata huvipakkuvate uuritavate tunnuste parameetreid nagu keskmised, osakaalud, ja kogusummad. Tänapäevased meetodid kasutavad uuritavate tunnuste kõrval ka teisi, nimelt abitunnuseid. Abitunnused sisaldavad teavet üldkogumi (või äärmisel juhul valimi) elementide kohta, mida kasutades saame oluliselt parandada hinnangute täpsust ja/või andmekogumise kulgu. Näiteks on Eesti rahvastikuregistris teavet isikute vanuse, soo, hariduse ja perekonna seisu kohta, mida saab ära kasutada kõikides valikuuringu etappides.

Üha rohkem tähelepanu koguv lähenemine on kohanduvad disainid (ingl. k. *responsive designs*), kus sekkutakse andmete kogumise protsessi eesmärgiga saada esinduslik vastanute hulk. Varasemalt on andmete kogumine lähtunud peamiselt kõrge vastamismäära saavutamisest. Kohanduvates disainides kasutatakse abitunnuseid.

Kohanduvas disainis võib näiteks kasutada tasakaalu mõistet ja tasakaalu indeksit andmete kogumise protsessis. Käesolevas töös on vastanute hulga tasakaalu all mõeldud abitunnuste keskmiste võrdumist vastanute hulgas ja valimis. Vastanute hulga tasakaalu mõõdame indeksiga IMB – skalaar, mis võrdleb abitunnuste keskmisi vastanute hulgas ja valimis.

Teaduslikus kirjanduses on mitmeid praktilisi uurimusi, kus näidatakse kohanduvate disainide positiivseid mõjusid, st. mitte-vastanutest tingitud hinnangu nihke vähendamisel. Antud dissertatsioonis esitame teoreetilised tulemused, kus näitame, et vastanute hulga tasakaalustamine indeksi IMB järgi vähendab suure nihke riski.

Kasulikku abiinformatsiooni saab leida ka hilises uuringu faasis, näiteks võivad abiinformatsiooni allikaks olla ka teised valikuuringud. Sageli viiakse läbi uuringuid samaaegselt ning tihti on neis uurimise all ühiseid tunnuseid, aga kuna neid viiakse ellu üksteisest sõltumatult, siis ühiste tunnuste hinnangud üldiselt erinevad. Tulemustes kooskõla saavutamine on dissertatsiooni üheks uurimisprobleemiks.

Antud dissertatsiooni eesmärk on panustada valikuuringute teooriasse kasutades abiinformatsiooni. Abiinformatsiooni kasutatakse andmete kogumise etapis, et saada tasakaalustatud vastanute hulk, ja hindamise etapis hinnangute parandamiseks kalibreerimise teel. Ühtlasi käsitletakse juhtu, kus uuritavat tunnust ennast saab käsitleda abiinformatsioonina, kui see pärineb välisest allikast.

Dissertatsiooni eesmärgid on detailselt järgmised:

- Eeldades, et uuringus on arvestatav kadu, kasutatakse andmete kogumisel meetodeid, mis parandavad vastanute hulga tasakaalu valitud abiinformatsiooni suhtes. Eesmärk on esitada teoreetilisi tõendeid, et pingutused vastanute hulga tasakaalustamisel parandavad hinnangute täpsust hindamise etapis. Eelkõige peame silmas kaost tingitud nihke vähendamist (Artikkel I);

- Otsides vastust dilemmale, et kas pigem kasutada abiinformatsiooni ainult hindamise etapil või hoopis panustada vastanute hulga tasakaalustamisele andmete kogumise etapil (või mõlemat), uurime erinevate abiinformatsiooni vektorite kasutamist. Eeldame, et hindamise etapil on meil saadaval rohkem abiinformatsiooni kui andmete kogumise etapil. Eesmärk on esitada kalibreeritud hinnang viisil kus lisatud abitunnuste mõju on otseselt eristatav (Artikkel II);
- Eeldades, et uuritava tunnuse kohta on teavet kahest andmeallikast (nt. muud uuringud või registrid), milles aga osakogumiteks (ehk üldkogumi gruppideks) jaotamine oli mõlemas allikas erineva tunnuse järgi. Käes-olevas uuringus huvitavad meid aga nende kahe tunnuse järgi ristklassifitseeritud osakogumid. Varem teadaolevat infot saab ära kasutada ning käes-olevas töös tuletame valemid ristklassifitseeritud osakogumite koos-kõlaliseks hindamiseks väliste andmeallikate olemasolu korral (Artikkel III);
- Illustreerida ja kinnitada kõikide teoreetilisi tulemusi simulatsioonides.

Dissertatsioon koosneb kolmest artiklist ja nende loetelu on esitatud töö alguses. Läbivalt eeldame, et järgime disainipõhist lähenemist ehk üldkogum on lõplik hulk, uuritavate tunnuste väärtused on fikseeritud ja juhuslikkus tuleneb viisist, kuidas kaasatakse üldkogumi elemente valimisse ehk valiku disainist. Järgnevalt on esitatud lühikokkuvõtted artiklite kaupa.

I artikkel

Antud artikkel on ühine artikkel koos professor Carl-Erik Särndali ja dotsent Imbi Traadiga, kus minu ülesanne oli läbi viia põhjalikud simulatsioonid. Lisaks osalesin teoreetilistes diskussioonides ja panustasin artikli valmimisse.

Artiklis eeldame, et uuringu käigus on tekkinud kadu ja uuritavate tunnuste väärtused on teada vaid vastanute hulgas. Seega pole võimalik kasutada hinnanguid, mis eeldavad kogu valimi vastamist. Üldiselt kasutatakse kaoga hindamisel kalibreeritud hinnanguid. Küsimuseks on, kui palju need erinevad kogu valimi pealt leitud hinnangutest (kui see oleks leitav)? Võtame aluseks kaost tingitud kalibreeritud hinnangu erinevuse kogu valimi pealt leitud nihketa hinnangust, tähistame selle vahe $\boldsymbol{\Delta}_r$, ja uurime tasakaalu indeksi IMB mõju $\boldsymbol{\Delta}_r$-le.

Artikli peamisteks tulemusteks on tasakaalu indeksi IMB ja $\boldsymbol{\Delta}_r$ vahelise seose tuletamine kahel erijuhul:

Tulemus 1. Esimesel juhul eeldatakse, et abiinformatsiooni vektor on indikaator vektor ehk grupivektor, mis määratleb millisesse gruppi valimielement kuulub. Sellisel juhul lihtsustuvad hinnangute valemid ning tasakaalu mõõdetakse selle järgi, kui palju vastamismäärad gruppides erinevad üldisest vastamismäärast.

Töös on näidatud, et $\boldsymbol{\Delta}_r$ tinglik keskväärtus on teatud tingimustel 0 ja tinglik dispersioon on seotud tasakaalu indeksiga IMB. Tasakaalu indeksi IMB vähenedes väheneb ka $\boldsymbol{\Delta}_r$ tinglik dispersioon ehk meil on väiksem tõenäosus saada suur erinevus kalibreeritud hinnangu ja kogu valimi hinnangu vahel. Osutub, et $\boldsymbol{\Delta}_r$ tinglik dispersioon sõltub ka uuritava tunnuse ja grupitunnuse seosest. Mida paremini grupeerimine seletab uuritavat tunnust, seda väiksem on $\boldsymbol{\Delta}_r$ tinglik dispersioon.

Tulemus 2. Teise tulemuse tuletamisel on hetkeks disainipõhise lähenemise asemel kasutatud mudeli põhist lähenemist ehk uuritavad tunnuse väärtused on juhuslikud ja juhuslikkust kirjeldab eeldatav mudel. Sellisel juhul saame loobuda grupivektori eeldusest, järeldused on üsna sarnased Tulemusega 1.

Mõlemaid tulemusi valideeriti simulatsioonidega. Joonis 3.1 illustreerib Tulemus 1 juhtu.

II artikkel

Artiklis käsitleme juhtu, kus andmete kogumist jälgitakse (monitooritakse) tasakaalu indeksi abil. Täpsemini, teatud ajahetkedel andmete kogumises arvutatakse tasakaalu indeks ja edasises andmekogumisprotsessis külastatakse vaid neid objekte, kes suurendavad vastanute hulga tasakaalu.

Seni on kohanduvate disainide uurimisel kasutatud enamasti üht ja sama abiinformatsiooni vektorit andmete kogumise ja hindamise etapis. Artiklis II eeldame, et hindamise etapis on meil kasutada rohkem abitunnuseid kui andmete kogumise ja monitoorimise etapis. Artiklis toodud uudsetest tulemuste eristame abitunnused kogu vektoris vastavalt monitoorimises kasutatud abitunnusteks ja lisatud abitunnusteks. Tuletame kalibreerimiskaalud nii, et eelnevalt eraldatud vektorite mõju on selgelt näha. Ilmutatud kujul kaalude abil esitame kalibreeritud hinnangu valemi. Tulemus heidab valgust huvitavatele erijuhtudele, nagu täieliku tasakaalu juht, sõltuvuse ja sõltumatuse juhud lisatud tunnuste ja monitoorimises kasutatavate tunnuste vahel.

Simulatsioonides käsitleme kahte huvitavat juhtumit. Esimesel juhul andmete kogumist ei monitoorita, keskendutakse hoopis lisa abitunnuste leidmisele, mida kasutatakse siis hindamise etapis hinnangute parandamiseks. Teisel juhul keskendutakse andmete kogumisele, lisa abitunnuseid hindamise etapis ei kaasata, kalibreerimis hinnangute leidmisel kasutatakse samu abitunnuseid, mida monitoorimisel.

Simulatsioonide tulemused näitavad, et kui lisatud abitunnused on tugevalt korelleeritud uuritava tunnusega, siis saame esimesel juhul keskmiselt väiksema nihkega hinnangud võrreldes teise juhuga. Kui aga lisatud abitunnused on nõrgalt korelleeritud uuritava tunnusega, siis saame paremad tulemused teisel juhul. Simulatsioonide ülesehitus lubas hinnata ka olukorda, kus tehakse mõlemaid tegevusi ehk monitooritakse vastanute hulga moodustumist ja hangitakse lisa abitunnuseid hindamise etapiks. Sellisel juhul saadi paremad tulemused võrreldes eelmise kahe juhuga. Ühtlasi oli näha, et monitoorimine parandas tasakaalu ka lisatud abitunnuste suhtes.

III artikkel

Eeldame, et meil on ligipääs kahele infoallikale, kus osakogumiteks jagamine on toimunud erinevate tunnuste järgi. Uues valikuuringus on kogutud andmestikus olemas mõlemad osakogumeid moodustavad tunnused ja moodustame ristklassifitseeritud osakogumid. Eelnevatest allikatest teadaoleva info paigutame marginaalidesse ja nõuame, et osakogumites leitud hinnangud ühilduksid vastava rea- või veeru marginaalidega. Kooskõla saavutamiseks kasutame AC-kalibreerimise meetodit (A - *auxiliary*, C - *common*), kus pikendame abiinformatsiooni vektorit suurustega, millega soovime kooskõla saada, ja korduv-kaalumise meetodit (*repeated weighting* - RW), kus juba olemasolevad kalibreerimiskaalud (A-informatsiooni abil) kalibreeritakse ümber C-informatsiooni abil, nii et summeerimiskitsendused oleksid rahuldatud.

Töös on esitatud valemid, kus uuritavaid tunnuseid võib olla mitu ja need on esitatud vektorina. Kõik ristklassifitseeritud kooskõlalised osakogumite hinnangud on leitavad ühe suure maatriksina.

Simulatsiooniülesandes testiti tuletatud valemeid ja võrreldi AC-kalibreerimise ja RW meetodeid. Selleks tekitati tehislik üldkogum Eesti Leibkonna Uuringu andmete põhjal. Kolme tunnuse abil moodustati kahemõõtmeline ristklassidega tabel, kus tuli kokku 12 osakogumit. Meetodite võrdlemiseks ja headuse hindamiseks kasutati baastasemena Horvitz-Thompsoni hinnanguid. Simulatsioonides käsitleti kahte juhtu, kus ühes kasutati täpseid marginaalsummasid, teises hinnati neid sõltumatust valimist.

AC-kalibreerimise ja RW meetodiga saavutati kooskõla teadaolevate suurustega teistest allikatest. Simulatsioonide tulemustest võib järeldada, et AC-kalibreeritud ja RW hinnangud on omavahel võrrelduna üsna sarnased, aga mõlemad olid väiksema standardhälbega kui HT hinnangud.

CURRICULUM VITAE

Kaur Lumiste

Address:	Männiku, Illi küla, Nõo vald, Tartumaa, Estonia
Telephone:	+372 53 991 695
E-mail:	kaur.lumiste@eesti.ee
Nationality:	Estonian
Date of Birth:	28th June 1985

Education

2011 - …	**PhD studies in Mathematical Statistics,** University of Tartu, Estonia Research topic: Improving accuracy of survey estimators by using auxiliary information in data collection and estimation stages
2008 - 2010	**MSc Mathematical Statistics**, University of Tartu, Estonia
2005 - 2008	**BSc Mathematical Statistics**, minor in Economics, University of Tartu, Estonia
1995 - 2004	Pärnu Co-Educational Gymnasium (awarded a silver medal for excellent curricular results)

Professional career

June '15 - …	**Titanium Systems**, data analyst
Mar ´12 – May'15	**University of Tartu, European Social Survey (ESS) Research Team**, analyst-consultant (Feb '14 – May '15 on paternity leave)
Dec ´10 – Feb ´12	**Estonian-Swedish Mental Health and Suicidology Institute (ERSI)**, statistician-consultant
Dec ´08 – June ´10	**BIGBANK AS, Management accounting department**, analyst

List of significant publications

Särndal, C.-E., Traat, I., **Lumiste, K.** (2017). Balancing survey response and reducing bias: Concepts and methods for estimation in times of high survey nonresponse. *Mathematical Population Studies*, 1–36 (to appear).

Särndal, C.-E., **Lumiste, K.**, Traat, I. (2016). Reducing the response imbalance: Is the accuracy of the survey estimates improved? *Survey Methodology, 42(2)*, 219–238 .

Yur'yev, A., Yur'yeva, L., Värnik, P., **Lumiste, K.**, Värnik, A. (2015). Complex Impact of Risk and Protective Factors on Suicide Mortality: A Study of Ukrainian General Population. *Archives of Suicide Research*, *19(2)*, 249–59.

Yur'yev, A., Värnik, P., Sisask, M., Leppik, L., **Lumiste, K.**, Värnik, A. (2013). Some aspects of social exclusion: Do they influence suicide mortality? *International Journal of Social Psychiatry, 59(3),* 232–238.

Ainsaar, M., Lilleoja, L., **Lumiste, K.**, Roots, A. (2013). *ESS Mixed Mode Experiment Results in Estonia (CAWI and CAPI Mode Sequential Design)*. Tartu: University of Tartu

Lumiste, K. (2011). Consistent Estimation of Cross-Classified Domains. *Statistics in Transition: new series*, *12*(2), 253–264.

Professional development

Jan-March 2016 **Statistical Learning**, Stanford University online course, USA, (3 months)

July 2016 **The 10th Tartu Conference on Multivariate Statistics**, Tartu, Estonia (4 days)

August 2015 **Baltic-Nordic Conference on Survey Sampling**, Helsinki, Finland (5 days)

August 2014 **Workshop of Baltic-Nordic-Ukrainian (BNU) Network on Survey Statistics 2014**, Tallinn, Estonia (5 days)

July 2013 **The 5th Annual Conference of the European Survey Research Association**, Ljubljana, Slovenia (5 days)

June 2013 **Workshop of BNU Network on Survey Statistics 2013**, Minsk, Belarus (5 days)

November 2012 **International Conference on European Social Survey**, Nicosa, Cyprus (3 days)

August 2012 **Workshop of Baltic-Nordic-Ukrainian Network on Survey Statistics 2012**, Valmiera, Latvia (6 days)

March - Dec 2012 **Reserve Officers' Basic Course**, Võru, Estonia (4 x 2 weeks)

June 2011 **Baltic-Nordic Conference on Survey Sampling**, Norrfälsviken, Sweden (5 days)

February 2011 **Strategies in comparative analysis: multi-level, multi-group and dummy approaches**, Barcelona, Spain (2 days)

Other administrative and professional activities

2013–2015 Council of institute of mathematical statistics, appointed member.

Oct. 2013 Public seminar on Survey methodology, Tartu, head organiser.

2012–2015 Researcher in SSHSS11196T "European Social Survey (ESS)", supervised by Mare Ainsaar, University of Tartu, Faculty of Social Sciences and Education, Institute of Sociology and Social Policy.

2011–2014 Researcher in ETF8789 "Consistency of domain estimators in case of multiple data sources", supervised by Imbi Traat, University of Tartu, Faculty of Mathematics and Computer Science.

2009–2010 Researcher in ETF7042 "Estimation in complex samples", supervised by Imbi Traat, University of Tartu, Faculty of Mathematics and Computer Science.

Public and social activities

2013 - … **International Association of Survey Statisticians (IASS)**, member

2012 - … **Estonian Statistical Society**, member

2011 -... **Estonian Defence League (Eesti Kaitseliit)**, active member

Additional information and distinctions

2017	Received Estonian League of Defence Medal of merits, III class
2011	Letter of Recognition from the National Science Contest for university level students
2004 - 2005	Military service 8 months, Pärnu Infantry Battalion, corporal’s rank for exemplary service
2004	Young Scientist’s badge laureate for scientific work

Qualifications

Languages	Estonian (native speaker)
	English (fluent in both speech and writing)
	Russian (basic in speech and writing)
	Finnish (basic in speech)
	Spanish (basic in speech)

CURRICULUM VITAE

Kaur Lumiste

Aadress	Männiku, Illi küla, Nõo vald, Tartumaa, Estonia
Telefon	+372 53 991 695
E-post	kaur.lumiste@eesti.ee
Rahvus	Eesti
Sünniaeg	28. juuni 1985

Haridustee

2011 - … **doktorantuur, matemaatiline statistika**, Tartu Ülikool, Eesti,

Uurimisteema: Valikuuringute hinnangute täpsuse tõstmine abiinformatsiooni kasutamisega andmekogumise ja hindamise etappides;

2008 - 2010 **magistrikraad, matemaatiline statistika**, Tartu Ülikool, Eesti;

2005 - 2008 **bakalaureus, matemaatiline statistika** (kõrvaleriala – majandus), Tartu Ülikool, Eesti;

1995 - 2004 Pärnu Ühisgümnaasium (lõpetatud hõbemedaliga).

Töökogemus

juuni ’15–… **Titanium Systems**, andmeanalüütik;

märts´12–mai’15 **Tartu Ülikool, Euroopa sotsiaaluuringu (ESS) uurimisrühm**, analüütik-konsultant (veebr. ’14 – mai ’15 olin isapuhkusel);

dets.´10–veebr.´12 **Eesti-Rootsi Vaimse Tervise ja Suitsidoloogia Instituut**, statistik-konsultant;

dets.´08–juuni´10 **BIGBANK AS, juhtimisarvestuse osakond**, analüütik.

Publikatsioonid

Särndal, C.-E., Traat, I., **Lumiste, K.** (2017). Balancing survey response and reducing bias: Concepts and methods for estimation in times of high survey nonresponse. *Mathematical Population Studies*, 1–36 (ilmumas).

Särndal, C.-E., **Lumiste, K.**, Traat, I. (2016). Reducing the response imbalance: Is the accuracy of the survey estimates improved? *Survey Methodology, 42(2)*, 219–238 .

Yur’yev, A., Yur'yeva, L., Värnik, P., **Lumiste, K.**, Värnik, A. (2015). Complex Impact of Risk and Protective Factors on Suicide Mortality: A Study of Ukrainian General Population. *Archives of Suicide Research*, *19(2)*, 249–59.

Yur’yev, A., Värnik, P., Sisask, M., Leppik, L., **Lumiste, K.**, Värnik, A. (2013). Some aspects of social exclusion: Do they influence suicide mortality? *International Journal of Social Psychiatry, 59(3),* 232–238.

Ainsaar, M., Lilleoja, L., **Lumiste, K.**, Roots, A. (2013). *ESS Mixed Mode Experiment Results in Estonia (CAWI and CAPI Mode Sequential Design)*. Tartu: University of Tartu

Lumiste, K. (2011). Consistent Estimation of Cross-Classified Domains. *Statistics in Transition: new series*, *12*(2), 253–264.

Erialased koolitused, konverentsid, seminarid

jaan.–märts 2016 **Statistical Learning**, Stanford Ülikooli veebikursus, USA, (3 kuud);

juuli 2016 **The 10th Tartu Conference on Multivariate Statistics**, Tartu, Eesti (4 päeva);

august 2015 **Baltic-Nordic Conference on Survey Sampling**, Helsinki, Soome (5 päeva);

august 2014 **Workshop of Baltic-Nordic-Ukrainian (BNU) Network on Survey Statistics 2014**, Tallinn, Eesti (5 päeva);

juuli 2013 **The 5th Annual Conference of the European Survey Research Association**, Ljubljana, Slovenia (5 päeva);

juuni 2013 **Workshop of BNU Network on Survey Statistics 2013**, Minsk, Valgevene (5 päeva);

november 2012 **International Conference on European Social Survey**, Nikosa, Küpros (3 päeva);

august 2012 **Workshop of Baltic-Nordic-Ukrainian Network on Survey Statistics 2012**, Valmiera, Läti (6 päeva);

märts – dets. 2012 **Vabatahtlike reservohvitseride kursus** , Võru, Eesti (4 x 2 nädalat);

juuni 2011 **Baltic-Nordic Conference on Survey Sampling**, Norrfälsviken, Rootsi (5 päeva);

veebruar 2011 **Strategies in comparative analysis: multi-level, multi-group and dummy approaches**, Barcelona, Hispaania (2 päeva).

Muu organisatsiooniline ja erialane tegevus

2013–2015 Matemaatilise statistika instituudi nõukogu liige.

Okt. 2013 Avalik küsitlusuuringute metodoloogia seminar, Tartu, peakorraldaja.

2012–2015 Põhitäitja projektis SSHSS11196T "Eesti osalemine Euroopa Sotsiaaluuringus", vastutav täitja Mare Ainsaar, Tartu Ülikool, Sotsiaal- ja haridusteaduskond, Sotsioloogia ja sotsiaalpoliitika instituut.

2011–2014 Põhitäitja projektis ETF8789 "Osakogumite hinnangute kooskõla saavutamine andmeallikate paljususe korral", vastutav täitja Imbi Traat, Tartu Ülikool, Matemaatika-informaatikateaduskond.

2009–2010 Põhitäitja projektis ETF7042 "Hindamine keeruliste valimite korral", vastutav täitja Imbi Traat, Tartu Ülikool, Matemaatika-informaatikateaduskond.

Ühiskondlik tegevus

2013 - … **International Association of Survey Statisticians (IASS)**, liige;

2012 - … **Eesti statistikaselts**, liige;

2011 -... **Eesti Kaitseliit**, aktiivne liige.

Lisainformatsioon ja autasud

2017	Omistati Eesti Kaitseliidu teenetemärgi III klass;
2011	Tänukiri, Eesti üliõpilaste teadustööde konkursil;
2004 - 2005	Ajateenistus 8 kuud, Pärnu Üksik-jalaväepataljonis, kaprali auaste eeskujuliku teenistuse eest;
2004	Noore teaduri märk teadustöö eest.

Oskused

keeled	eesti keel (emakeel)
	inglise keel (kirjalik ja suuline suhtlemine väga hea)
	vene keel (esmase suhtluse ja kirjalikul tasandil)
	soome keel (esmase suhtluse tasandil)
	hispaania keel (esmase suhtluse tasandil)

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. **Mati Heinloo.** The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu, 1991, 23 p.
2. **Boris Komrakov.** Primitive actions and the Sophus Lie problem. Tartu, 1991, 14 p.
3. **Jaak Heinloo.** Phenomenological (continuum) theory of turbulence. Tartu, 1992, 47 p.
4. **Ants Tauts.** Infinite formulae in intuitionistic logic of higher order. Tartu, 1992, 15 p.
5. **Tarmo Soomere.** Kinetic theory of Rossby waves. Tartu, 1992, 32 p.
6. **Jüri Majak.** Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu, 1992, 32 p.
7. **Ants Aasma.** Matrix transformations of summability and absolute summability fields of matrix methods. Tartu, 1993, 32 p.
8. **Helle Hein.** Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu, 1993, 28 p.
9. **Toomas Kiho.** Study of optimality of iterated Lavrentiev method and its generalizations. Tartu, 1994, 23 p.
10. **Arne Kokk.** Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu, 1995, 165 p.
11. **Toomas Lepikult.** Automated calculation of dynamically loaded rigid-plastic structures. Tartu, 1995, 93 p, (in Russian).
12. **Sander Hannus.** Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74 p, (in Russian).
13. **Sergei Tupailo.** Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu, 1996, 134 p.
14. **Enno Saks.** Analysis and optimization of elastic-plastic shafts in torsion. Tartu, 1996, 96 p.
15. **Valdis Laan.** Pullbacks and flatness properties of acts. Tartu, 1999, 90 p.
16. **Märt Põldvere.** Subspaces of Banach spaces having Phelps' uniqueness property. Tartu, 1999, 74 p.
17. **Jelena Ausekle.** Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu, 1999, 72 p.
18. **Krista Fischer.** Structural mean models for analyzing the effect of compliance in clinical trials. Tartu, 1999, 124 p.
19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
20. **Jüri Lember.** Consistency of empirical k-centres. Tartu, 1999, 148 p.
21. **Ella Puman.** Optimization of plastic conical shells. Tartu, 2000, 102 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.

23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
25. **Maria Zeltser.** Investigation of double sequence spaces by soft and hard analitical methods. Tartu, 2001, 154 p.
26. **Ernst Tungel.** Optimization of plastic spherical shells. Tartu, 2001, 90 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 p.
28. **Rainis Haller.** *M(r,s)*-inequalities. Tartu, 2002, 78 p.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
30. Töö kaitsmata.
31. **Mart Abel.** Structure of Gelfand-Mazur algebras. Tartu, 2003. 94 p.
32. **Vladimir Kuchmei.** Affine completeness of some ockham algebras. Tartu, 2003. 100 p.
33. **Olga Dunajeva.** Asymptotic matrix methods in statistical inference problems. Tartu 2003. 78 p.
34. **Mare Tarang.** Stability of the spline collocation method for volterra integro-differential equations. Tartu 2004. 90 p.
35. **Tatjana Nahtman.** Permutation invariance and reparameterizations in linear models. Tartu 2004. 91 p.
36. **Märt Möls.** Linear mixed models with equivalent predictors. Tartu 2004. 70 p.
37. **Kristiina Hakk.** Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 137 p.
38. **Meelis Käärik.** Fitting sets to probability distributions. Tartu 2005, 90 p.
39. **Inga Parts.** Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140 p.
40. **Natalia Saealle.** Convergence and summability with speed of functional series. Tartu 2005, 91 p.
41. **Tanel Kaart.** The reliability of linear mixed models in genetic studies. Tartu 2006, 124 p.
42. **Kadre Torn.** Shear and bending response of inelastic structures to dynamic load. Tartu 2006, 142 p.
43. **Kristel Mikkor.** Uniform factorisation for compact subsets of Banach spaces of operators. Tartu 2006, 72 p.
44. **Darja Saveljeva.** Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 117 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
46. **Annely Mürk.** Optimization of inelastic plates with cracks. Tartu 2006. 137 p.
47. **Annemai Raidjõe.** Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97 p.
48. **Olga Panova.** Real Gelfand-Mazur algebras. Tartu 2006, 82 p.

49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
50. **Margus Pihlak.** Approximation of multivariate distribution functions. Tartu 2007, 82 p.
51. **Ene Käärik.** Handling dropouts in repeated measurements using copulas. Tartu 2007, 99 p.
52. **Artur Sepp.** Affine models in mathematical finance: an analytical approach. Tartu 2007, 147 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
54. **Kaja Sõstra.** Restriction estimator for domains. Tartu 2007, 104 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
57. **Evely Leetma.** Solution of smoothing problems with obstacles. Tartu 2009, 81 p.
58. **Ants Kaasik.** Estimating ruin probabilities in the Cramér-Lundberg model with heavy-tailed claims. Tartu 2009, 139 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
60. **Indrek Zolk.** The commuting bounded approximation property of Banach spaces. Tartu 2010, 107 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
63. **Marek Kolk.** Piecewise Polynomial Collocation for Volterra Integral Equations with Singularities. Tartu 2010, 134 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
65. **Larissa Roots.** Free vibrations of stepped cylindrical shells containing cracks. Tartu 2010, 94 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
68. **Olga Liivapuu.** Graded q-differential algebras and algebraic models in noncommutative geometry. Tartu 2011, 112 p.
69. **Aleksei Lissitsin.** Convex approximation properties of Banach spaces. Tartu 2011, 107 p.
70. **Lauri Tart.** Morita equivalence of partially ordered semigroups. Tartu 2011, 101 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.

72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.
74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
75. **Nadežda Bazunova.** Differential calculus $d^3 = 0$ on binary and ternary associative algebras. Tartu 2011, 99 p.
76. **Natalja Lepik.** Estimation of domains under restrictions built upon generalized regression and synthetic estimators. Tartu 2011, 133 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
80. **Marje Johanson.** $M(r, s)$-ideals of compact operators. Tartu 2012, 103 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
82. **Vitali Retšnoi.** Vector fields and Lie group representations. Tartu 2012, 108 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
85. **Erge Ideon**. Rational spline collocation for boundary value problems. Tartu, 2013, 111 p.
86. **Esta Kägo.** Natural vibrations of elastic stepped plates with cracks. Tartu, 2013, 114 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
88. **Boriss Vlassov.** Optimization of stepped plates in the case of smooth yield surfaces. Tartu, 2013, 104 p.
89. **Elina Safiulina.** Parallel and semiparallel space-like submanifolds of low dimension in pseudo-Euclidean space. Tartu, 2013, 85 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
93. **Kerli Orav-Puurand.** Central Part Interpolation Schemes for Weakly Singular Integral Equations. Tartu, 2014, 109 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.

95. **Kaido Lätt.** Singular fractional differential equations and cordial Volterra integral operators. Tartu, 2015, 93 p.
96. **Oleg Košik.** Categorical equivalence in algebra. Tartu, 2015, 84 p.
97. **Kati Ain.** Compactness and null sequences defined by ℓ_p spaces. Tartu, 2015, 90 p.
98. **Helle Hallik.** Rational spline histopolation. Tartu, 2015, 100 p.
99. **Johann Langemets.** Geometrical structure in diameter 2 Banach spaces. Tartu, 2015, 132 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
105. **Md Raknuzzaman.** Noncommutative Galois Extension Approach to Ternary Grassmann Algebra and Graded q-Differential Algebra. Tartu, 2016, 110 p.
106. **Alexander Liyvapuu.** Natural vibrations of elastic stepped arches with cracks. Tartu, 2016, 110 p.
107. **Julia Polikarpus.** Elastic plastic analysis and optimization of axisymmetric plates. Tartu, 2016, 114 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.
113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
115. **Tiina Kraav.** Stability of elastic stepped beams with cracks. Tartu, 2017, 126 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.

117. **Silja Veidenberg.** Lifting bounded approximation properties from Banach spaces to their dual spaces. Tartu, 2017, 112 p.
118. **Liivika Tee.** Stochastic Chain-Ladder Methods in Non-Life Insurance. Tartu, 2017, 110 p.
119. **Ülo Reimaa.** Non-unital Morita equivalence in a bicategorical setting. Tartu, 2017, 86 p.
120. **Rauni Lillemets.** Generating Systems of Sets and Sequences. Tartu, 2017, 181 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.