VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Vilma NEKRAŠAITĖ-LIEGĖ

SMALL AREA ESTIMATION

DOCTORAL DISSERTATION

PHYSICAL SCIENCES,
MATHEMATICS (01P)

Vilnius LEIDYKLA TECHNIKA 2012

Doctoral dissertation was prepared at Vilnius Gediminas Technical University
in 2007–2012.

Scientific Supervisor

Assoc Prof Dr Marijus RADAVIČIUS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P).

http://leidykla.vgtu.lt
VGTU leidyklos TECHNIKA 2099-M mokslo literatūros knyga

ISBN 978-609-457-410-8

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Vilma NEKRAŠAITĖ-LIEGĖ

MAŽŲ SRIČIŲ VERTINIMAS

DAKTARO DISERTACIJA

FIZINIAI MOKSLAI,
MATEMATIKA (01P)

Vilnius VGTU LEIDYKLA TECHNIKA 2012

Disertacija rengta 2007–2012 metais Vilniaus Gedimino technikos universitete.

Mokslinis vadovas

doc. dr. Marijus RADAVIČIUS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

Abstract

In the dissertation special problems that may be encountered in finding optimal estimation strategy for small area estimation, in particular, model diagnostics for small area models, constrained estimation, sample design selection, nonresponse adjustment and borrowing strength across both small areas and time are considered.

The estimation strategy is a combination of sampling design and estimation design. First of all several well known sample designs and population total estimators are constructed for small areas. The changes of these estimators are showed in the case of the use of various nonresponse adjustment methods. A simulation using a real population from Statistics Lithuania is done to investigate the performance of different types of estimation strategies when various problems occurs: small area, nonresponse. The different underlying models are examined for both design-based model assisted and model-based estimators. This study showed that the nonresponse has bigger negative effect for design-based estimators than for model-based estimators. Still, generally, the design-based model assisted estimator performs better than model-based estimator.

To improve estimation strategy a balance sample and model-based sample design are introduced. The model-based sample design is based on the historical data, which are used to construct superpopulation model before sample selection. The variance of prediction error is used to construct inclusion probabilities, thus the element with larger variance of prediction error have larger probability to be selected in to the sample. The simulation studies showed, that in many cases the use of model-based sample design reduce the accuracy measures.

Reziumė

Disertacijoje nagrinėmos problemos, iškylančios ieškant geriausios mažų sričių vertinimo strategijos. Ieškant geriausios mažų sričių vertinimo strategijos susiduriama su modelio parinkimo, imties plano ir įvertinio konstravimo, neatsakymų vertinimo ir papildomos informacijos panaudojimo problemomis.

Vertinimo strategija – tai imties plano ir įvertinio kombinacija. Visų pirma disertacijoje pateikiama gerai žinomų imties planų ir įvertinių išraiškos mažų sričių vertinimo atveju. Taip pat pateikiama kaip nagrinėjami įvertiniai pasikeičia jei neatsakymų vertinimui pasirenkami įvairūs persvėrimo ar duomenų įrašymo metodai. Modeliavimams panaudota reali populiacija gauta iš Lietuvos statistikos departamento. Pasinaudojant modeliavimu buvo tiriama imties dydžio mažoje srityje bei neatsakymų vertinimo metodų įtaka skirtingoms vertinimo strategijoms. Taip pat buvo tiriama pasirinkto modelio įtaka imties planu paremtiems ir modeliu pagrįstiems įvertiniams. Rezultatai parodė, jog neatsakymai labiau paveikia imties planu paremtų įvertinių savybes, tačiau daugeliu atveju imties planu paremti įvertiniai naudojantys modelį tik kaip pagalbinį įrankį yra pranašesni nei modeliu pagrįsti įvertiniai.

Siekiant pagerinti vertinimo strategiją taip pat buvo nagrinėjamos subalansuotos imtys ir modeliu pagrįsti imties planai. Pastarieji yra paremti istoriniais duomenimis, kurie yra naudojami modelio konstravimui prieš imties išrinkimą. Tokiuose planuose priklausymo imčiai tikimybės yra konstruojamos atsižvelgiant į modelio paklaidų dispersijas: elementai kurių modelio paklaidų dispersija yra didelė turi didesnę priklausymo imčiai tikimybę, nei tie elementai, kurių modelio paklaidų dispersija yra maža. Modeliavimo rezultatai parodė, kad daugeliu atveju naudojant modeliu pagrįstus imties planus įverčių poslinkis ir dispersija sumažėja.

Notations

U – finite population consisting of N units;
N – population size;
$U^{(d)}$ – population domain consisting of $N^{(d)}$ units;
D – number of domains;
$N^{(d)}$ – a domain size in d domain;
t – time;
$y_k(t)$ – a value of a study variable y for element k in time t;
$\mathbf{x}_k(t) = \{x_{1,k}(t), x_{2,k}(t), \ldots, x_{J,k}(t)\}$ – the values of J auxiliary variables in time t;
$\mathbf{q}_k = \{q_k^{(1)}, q_k^{(2)}, \ldots, q_k^{(D)}\}$ – the domain indicators;
$y_k^{(d)}(t)$ – a value of domain variable for element k in time t;
$TOT^{(d)}(t)$ – a domain total of a study variable in time t;
$\underline{\mathbf{S}}$ – a sampling vector;
$\mathbf{S}(t)$ – a sample;
$n(t)$ – a sample size in time t;
$s(t)$ – a sample set of $\mathbf{S}(t)$;
$U_{ns}(t)$ – a non sampled set of $\mathbf{S}(t)$;
$p(\mathbf{S}(t))$ – a sample design;
$\pi_k(t)$ – an inclusion probability for unit k in time t;
$\pi_{kl}(t)$ – an inclusion probability for units k and l in time t;

$w_k(t)$ – a sampling weight for unit k in time t;
$s^{(d)}(t)$ – a sample set in domain;
$n^{(d)}(t)$ – a sample size in d domain;
$r(t)$ – a response set of $\mathbf{S}(t)$;
$s_{nr}(t)$ – a non response set of $\mathbf{S}(t)$;
$\hat{y}_k(t)$ – a predicted value of a study variable for unit k in time t;
$\underline{\hat{\theta}}(t)$ – an estimator of parameter θ;
$\hat{\theta}(t)$ – an estimate of parameter θ;
$\underline{\widehat{TOT}}^{(d)}(t)$ – an estimator of domain total in time t;
$\widehat{TOT}^{(d)}(t)$ – an estimate of domain total in time t;
$BIAS(\underline{\hat{\theta}}(t))$ – a bias of an estimator $\underline{\hat{\theta}}(t)$;
$E_p(\underline{\hat{\theta}}(t))$ – a mean of an estimator $\underline{\hat{\theta}}(t)$ under the sample design;
$var_p(\underline{\hat{\theta}}(t))$ – a variance of an estimator $\underline{\hat{\theta}}(t)$ under the sample design;
$ARB(\widehat{TOT}^{(d)}(t))$ – an absolute relative bias of $\widehat{TOT}^{(d)}(t)$;
$RRMSE(\widehat{TOT}^{(d)}(t))$ – a relative root means square error $\widehat{TOT}^{(d)}(t)$;
$MARB(\widehat{TOT}^{(d)}(t))$ – a mean of absolute relative bias of $\widehat{TOT}^{(d)}(t)$;
$MRRMSE(\widehat{TOT}^{(d)}(t))$ – a mean of relative root means square error $\widehat{TOT}^{(d)}(t)$.

Contents

Introduction

Scientific problem

A general task of survey sampling is formulated such way: a finite population is given and it is needed to estimate some parameter (for example, population total) for study variable, when the information about the values of a study variable is know not for all elements. An estimation strategy is a pair of sample design and estimator. It is searching the better strategies in one or other way. In this dissertation a strategies are analyzed not for all population, but for the small areas.

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. Such area estimation is becoming important in survey sampling due to a growing demand for reliable small area statistics from both public and private sectors. A lot of surveys are carried out not once, but from time to time, thus a huge dataset of auxiliary information is stored. The possibility to use this information for improvement of small area estimation is analyzed in this dissertation.

Also there is no survey, that there was no nonresponse. The estimates become biased if there is no nonresponse adjustment done. In this dissertation the effect of different nonresponse adjustments methods are analyzed in the case of small area estimation.

Topicality of the work

The nonresponse occurs in all surveys. It decreases the estimator's accuracy, thus several methods of nonresponse adjustments are developed (Rubin (1987), Särndal and Lundstrom(2005)). However these methods are used just for estimating the parameters of the whole population or large domain. To the best of our knowledge, the influence of nonresponse for small area estimation is not analyzed. Naturally it is desirable to examine this influence at least using simulation technique.

Nowadays, official statistics repeats the same surveys from year to year, so for most of the population elements it is possible to get information for the same variable in several time periods. It means that for many surveys individual data for some objects are known for at least one previous time point. We will call this panel-type data even when it is not a part of the design. Also, in some cases it is possible to use information collected from the other sources (tax offices, jobcenters, etc.). Such datasets of a large amount of auxiliary information might improve the quality of the estimation strategy as compared with a strategy based on the current sample alone. Ghosh and Rao (1994) have already applied this kind of information to construct area level model for estimating small area estimations. Naturally it is desirable to construct an unit level model and to use it not only in the estimation stage, but also for the sample design construction.

The survey sampling is quite a new science. The first results were developed after 1940. In Lithuania, the first work from survey sampling was published just after 1991. Today the survey sampling are widely used in all kind of the surveys, however treating a population as fixed still prevails in the theory of survey sampling. In the dissertation, an attempt to develop a superpopulation model which takes into account both dynamics and randomness of population elements as well as the finiteness of the population (and hence, a sampling design) is made.

Research object

The research object is properties of the estimating strategies for small area totals with the aim of their improvement. Thus the objects of interest are: sample designs, estimators and superpopulation models used in the small area estimation.

The aim and tasks of the dissertation

The aim of the dissertation is, using real survey data, to find the best population total in small areas estimating strategy, including nonresponse adjustment, among the set of available strategies, and to propose a methodology for the best strategy selection in real regular (repeated) surveys.

Let us state the following problems:

1. To review sample designs and estimators already used in small area estimation and to form a set of estimating strategies to be compared.
2. To review nonresponse adjustment methods and to examine what of these methods are best for small area estimation.
3. To propose a superpopulation model for a finite population with random and varying in time elements.
4. Research of optimal strategy for small area estimation in the case when the same variables are measured from time to time.

Research methods

The superpopulation model is based on panel data models. Also generalized regression (GREG), stratified samples, balanced sampling design and other survey sampling methods are applied. The main results are obtained by computer simulations based on a real survey data. All calculations are performed using SAS software.

Scientific novelty

1. The proposed superpopulation model based on panel data models enables one to represent a finite population with both random and varying

in time elements;

2. The effect (properties and etc.) of nonresponse adjustment methods in small area estimation were not studied.
3. A model-based sample design for repeated surveys is constructed and it is demonstrated using real survey data that in most of the cases it works better than the other sampling designs;
4. A methodology of best strategy selection for repeated surveys is proposed.

Practical value of the work results

A methodology of best strategy selection for small area estimation in repeated surveys of official statistics is proposed. The stress on small area estimation is made because of increasing demand of reliable statistics at lower geographic and statistical classification of economic activities in the European Community (NACE) levels, especially from local governments and from businesses, in order to make investment, marketing, and location decisions that depend on knowledge of local areas.

Propositions presented for defense

1. Donors methods for nonresponse adjustment in the case of small area estimation should not be used.
2. Panel data models can be used to represent randomness and time variability in real finite populations.
3. It was revealed that in the cases when there is auxiliary variable well correlated with study variable (correlation coefficient is more than 0.9) the best strategy is to use simple regression model as the assisted tool for GREG estimation.
4. It is demonstrated by computer simulations using real survey data that in the cases where a large amount of auxiliary information from the past is available the model-based sample design might be the best choice.

Approval of the work results

The main results are published in six articles and presented in the conferences and workshops:

1. Nekrašaitė-Liegė, V. Sumos vertinimas mažose srityse, Lietuvos matematikų draugijos konferencija, Kaunas, 2008 m. birželio 25–26 d.
2. Nekrašaitė-Liegė, V. Small area estimation in practice, Workshop on Survey Sampling Theory and Methodology, Kuressaare, Estonia, 2008 August 25–29 d.
3. Nekrašaitė-Liegė, V. Neatsakymų įtaka populiacijos sumos vertinimui, 12-oji Lietuvos jaunųjų mokslininkų konferencija, Vilnius, 2009 m. balandžio 16 d.
4. Nekrašaitė-Liegė, V. Mažų sričių vertinimas neatsakymų atveju, Lietuvos matematikų draugijos konferencija, Vilnius, 2009 m. birželio 18–19 d.
5. Nekrašaitė-Liegė, V. Persvėrimų ir įrašymo metodų palyginimas mažose srityse, Lietuvos matematikų draugijos konferencija, Šiauliai, 2010 m. birželio 17–18 d.
6. Nekrašaitė-Liegė, V., Radavičius, M., Rudys, T. Model-based design in small area estimation, 10-th International Vilnius Conference on Probability Theory and Mathematical Statistics, Vilnius, 2010-06-28 — 2010-07-02.
7. Nekrašaitė-Liegė, V. Nonresponse adjustment in SAE under different sampling designs, Workshop on Survey Sampling Theory and Methodology, Vilnius, 2010 rugpjūčio 23–27 d.
8. Nekrašaitė-Liegė, V. Some applications of panel data models in small area estimation, BaNoCoSS 2011, Norrfällsviken, Švedija, 2011 birželio 13–17 d.
9. Nekrašaitė-Liegė, V. Mažų sričių vertinimo strategijų palyginimas, Respublikinė jaunųjų mokslininkų konferencija „Fundamanetiniai tyrimai ir inovacijos mokslų sandūroje 2012”, Klaipėda, 2012 balandžio 20 d.
10. Nekrašaitė-Liegė, V. Mažų sričių vertinimo metodika naudojant panelinius duomenis, Lietuvos matematikų draugijos konferencija, Klaipėda, 2012 birželio 11–12 d.

11. Nekrašaitė-Liegė, V. Estimation strategy for small areas, a case study, Workshop on Survey Sampling Theory and Methodology, Valmiera, Latvija, 2012 rugpjūčio 24–28 d.

The scope of the scientific work

The scientific work layout consists of introduction, four chapters, conclusions, references and lists of authors publications. The total scope of the dissertation is 90 pages, 17 tables, 80 items of reference.

The first chapter is the literature overview, which presents the other authors results on the topic of the dissertation. All the results are divided into three separate parts: models in small area, nonresponse adjustment, estimation strategy. Thus, the results of this dissertation are new, because there is no literature that combines all these parts.

The main notations and definitions used in this dissertation are presented in the second chapter. Also this chapter is dedicated for the estimators, nonresponse adjustment methods and the main models which can be used in small area estimation.

The different searches of optimal strategy are presented in the third chapter. Here the balance sample and model-based sample design are introduced.

The simulation results, which show the performance of the different small area estimators in the case of nonresponse and comparison of different estimation strategies are presented in the fourth chapter.

1

Literature about small area estimation overview

1.1. Models in small area

As mentioned by Ghosh and Rao (1994) the term “small area” and “local are” 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.

Sample sizes for small areas are typically small because the overall sample size in a survey is usually determined to provide desired accuracy at a much higher level of aggregation. To provide desired accuracy at small area level, models are used.

Small area models may be broadly classified into two types: area level and unit level.

The basic area level model was developed by Fay and Herriot (1979). It has been extended to handle correlated sampling errors, spatial dependence of random small area effects, time series and cross-sectional data and others (see Ghosh and Rao (1994)).

Singh, Stukel and Pfeffermann (1998) made a comparison of frequentest and Bayesian measures of error, using analytical and empirical methods for the basic unit-level model.

The basic unit level model was developed by Battese, Harter and Fuller (1988). Various extensions of the basic unit level models have been studied. Stukel and Rao (1999) studied two-way nested error regression models which are appropriate for two-stage sampling within small areas. Following Kleffe and Rao (1992), Arora and Lahiri (1997) studied unit level models with random error variances. Kleffe and Rao (1992) assumed the existence of only mean and variance, without specifying a parametric distribution on variance. Datta, Day and Basawa (1999) extended the unit level model to the multivariate case following Fuller and Harter (1987). This extension leads to a multivariate nested error regression model. Moura and Holt (1999) allow some or all of the regression coefficients to be random and to depend on area level auxiliary variables, thus effectively integrating the use of unit level and area level covariates into a single model. Malec, Davis and Cao (1999) and Malec, Sedransk, Moriarity and LeClere (1997) studied the binary case, using logistic linear mixed models with random slopes to link the small areas.

All main models used in small area estimation are described by Ghosh and Rao (1994), Rao (1999), Rao (2003). Rao and Choudhry (1995) provided an overview of small area estimation in the context of business surveys. Still, the literature about the use of panel data model for small area estimation was not found.

1.2. Nonresponse adjustment

Nonresponse is unavoidable in surveys. It is classified as unit nonresponse, which occurs when, for a sample unit, all the survey variables are missing or when not enough usable information is available, and item nonresponse when, for a sample unit several, but not all survey variables are missing.

Weighting adjustment is a popular method for handling unit nonresponse in sample surveys. Groves, Dillman, Eltinge and Little (2002), Särndal, Lundstrom (2005) provided comprehensive overviews of nonresponse weighting adjustment (NWA) methods in survey sampling. There are two types of NWA: nonresponse propensity weighting (NPW) and nonresponse calibration weighting (NCW).

When the estimated response probability is directly used and no other adjustment is made, the method is called the direct NWA method (see Rosenbaum (1987)). Applications of the direct NWA method can be found in Ekholm and Laaksonen (1991), Folsom and Singh (2000), and Iannacchione (2003). Bethlehem (1988) and Fuller and An (1998) discuss the regression NWA method.

The second type of adjustment procedures, called nonresponse calibration weighting (NCW) can be seen as an extension of the calibration approach (Deville and Särndal (1992)) adapted to the context of unit nonresponse. The reader is referred to Lundström and Särndal (1999), Särndal, Lundstrom (2005) and Kott (2006) for a comprehensive overview of NPW and NWC.

The problem of variance estimation in the context of NPW has been recently studied by Kim and Kim (2007) and in context of NCW – by Haziza, Thompson and Yung (2010). Kim and Kim (2007) showed that the estimator using the estimated response probability is more efficient than the estimator using the true response probability when the parameters for response probabilities are estimated by the maximum likelihood method. Haziza, Thompson and Yung (2010) considered two jackknife variance estimators.

As for the item nonresponse, imputation can be used. The important practical problem of estimating the variance of an estimate computed from a data set in which some of the items are missing and values are assigned by imputation has been addressed in a number of different ways (e.g., see Rubin (1987) and Rao and Shao (1992)). The problem of variance estimation for a linear estimator in which missing values are assigned by a single hot deck imputation (a form of imputation that is widely used in practice) is studied by Brick, Kalton and Kim (2004).

Still all these papers, which are mentioned above examine nonresponse adjustment problems in the context of the whole population, but not in small areas or even domains. Just Brick, Jones, Kalton and Valliant (2005) demonstrate, that even if an imputation method gives almost unbiased estimates for the full population, estimates for domains may be very biased.

1.3. Estimation strategy

As Singh, Gambino and Mantel (1994) point out: “where possible, samples should be designed to produce small area estimates of adequate precision, and sample designs should be fashioned with this in mind. Auxiliary data should be used, where possible, to improve the precision of direct small area estimates.” One of the possible way to improve the precision might be the use of balanced sampling.

In the model-based framework, Royall (1976a) advocated the use of balanced sampling in order to reach the optimal strategy and to protect against misspecification of the model. (see also Royall (1976b), Royall and Pfeffermann (1982), Kott (1986), Cumberland and Royall (1988), Royall (1988)). Nedyalkova, Tille (2008) showed, that in the model-assisted and the model-based frameworks, a balancing sampling design with the Horvitz-Thompson estimator is often an optimal strategy. Indeed, when the sample is balanced, the variances of the Horvitz-Thompson estimators of the auxiliary variables are equal to zero. Under a linear model, the variance of the Horvitz-Thompson estimator of the interest variable will only depending on the residuals of the model.

For the whole population estimation the most efficient strategy always consists of using balanced sampling and calibration together (see the simulation in Deville and Tillé (2004)).

The main problem of using balance sample is to select such sample. The first method for selecting a random balanced sample were proposed by Yates (1946), but this method was rejective in the sense that it involved selecting samples randomly in the sample until a balanced enough sample was obtained.

Deville, Grosbras and Roth (1988) and Deville (1992) proposed multivariate methods for balanced sampling with equal inclusion probabilities. Hedayat and Majumdar (1995) have proposed the adaptation of an experimental design technique that would enable a balanced sampling design to be constructed. Again, this technique is restricted to equal inclusion probabilities. Finally, the cube method was proposed by Deville and Tillé (2004). It is an extension of the splitting method that was developed by Deville and Tillé (1998). 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. There is and some other method to get a balance sampling (see Fuller (2009)), but it is not so popular as cube method.

To program cube method is quite difficult, that is why free software programs are available. One of them is done by Chauvet and Tillé (2005). It is an SAS/IML implementation and is available on the University of Neuchatel Web site. In R language, the sampling package (Tillé and Matei (2007)) also allows us to use the cube method.

The disadvantage of cube method is that, it is balanced for whole population, but not for small areas or even domains. Falorsi and Righi (2008) studied a balanced sampling approach for multi-way stratification designs for small area estimation. The proposed sampling strategy is based on the use of both a balanced sampling selection technique (Deville and Tillé (2004)) and a GREG-type estimation (Lehtonen, Särndal and Veijanen (2003)). The study showed, that in some survey context, the proposed sampling strategy might define a too large overall sample size for assuring the prefixed bound of the direct domain estimates sampling errors. If the overall sample size is bounded by budget constraints, then the proposed sampling strategy with direct estimators may be not feasible.

The literature about model-based design was not found at all.

2

The elements of estimation strategy

2.1. Definitions and notations

2.1.1. Population, auxiliary information and study variables

Let us start with a common framework of finite population survey sampling. 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 component $y(t)$ defines the value of a study variable (variable of interest), 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. The values of these two components depends on time t, $t = 1, 2, ...$

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:

$$q_k^{(d)} = \begin{cases} 1, & \text{if } k \in U^{(d)}, \\ 0, & \text{otherwise,} \end{cases} \qquad \forall k \in U, \quad d = 1, \ldots, D. \tag{2.1}$$

In dealing with a domain, $U^{(d)}$, it is convenient to use the domain specific variable, $y^{(d)}(t)$ defined as $y_k^{(d)}(t) = y_k(t)$ if $k \in U^{(d)}$, and $y_k^{(d)}(t) = 0$ if $k \notin U^{(d)}$.

For every $k \in U$, values $q_k^{(d)}$, $d = 1, \dots, D$ construct a domain indicator column vector $\mathbf{q}_k = (q_k^{(1)}, q_k^{(2)}, \dots, q_k^{(D)})$. Hence all properties of unit k in time t are in vector $\mathbf{a}_k(t) = (\mathbf{x}_k(t), \mathbf{q}_k, y_k(t))$ of dimension $(J+D+1) \times 1$ and the properties of the population U in time t are

$$\mathbf{A}(t) = \begin{bmatrix} \mathbf{a}_1'(t) \\ \mathbf{a}_2'(t) \\ \vdots \\ \mathbf{a}_N'(t) \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1'(t) & \mathbf{q}_1' & y_1(t) \\ \mathbf{x}_2'(t) & \mathbf{q}_2' & y_2(t) \\ \vdots & \vdots & \vdots \\ \mathbf{x}_N'(t) & \mathbf{q}_N' & y_N(t) \end{bmatrix}. \tag{2.2}$$

The matrix $\mathbf{A}(t)$ of dimension $N \times (J+D+1)$ is called the data matrix in time t. Its row vectors $\mathbf{a}_k(t)$ correspond to units k and the column vectors correspond to the properties associated with the units. We denote

$$\begin{aligned} \mathbf{X}(t) &= (\mathbf{x}_1'(t), \mathbf{x}_2'(t), \dots, \mathbf{x}_N')(t), \\ \mathbf{Q} &= (\mathbf{q}_1', \mathbf{q}_2', \dots, \mathbf{q}_N'), \\ Y(t) &= (y_1(t), y_2(t), \dots, y_N(t)). \end{aligned} \tag{2.3}$$

Now the data matrix (2.2) can be written as $\mathbf{A}(t) = [\ \mathbf{X}(t) \quad \mathbf{Q} \quad Y(t)\ ]$.

2.1.2. Parameters of interest and sampling design

The parameter of interest is a domain total:

$$TOT^{(d)}(t) = \sum_{k \in U^{(d)}} y_k(t) = \sum_{k \in U} q_k^{(d)} y_k(t) = \sum_{k \in U} y_k^{(d)}(t), \quad d = 1, \dots, D. \tag{2.4}$$

To estimate $TOT^{(d)}(t)$, we need information about unknown variable $y(t)$. This information is collected by sampling. The sampling vector

$$\underline{\mathbf{S}} = (\underline{S}_1, \underline{S}_2, \dots, \underline{S}_N) \tag{2.5}$$

is a random vector whose elements $\underline{S}_k$ indicate the number of selections for k. The realization $\mathbf{S}(t) = (S_1(t), S_2(t), \dots, S_N(t))$ in time t is called a sample. Let $\mathcal{S}(t)$ be the set of all samples $\mathbf{S}(t)$. The sampling vector $\underline{\mathbf{S}}$ (and its realization $\mathbf{S}(t)$) define the sample set $\underline{s}$ (and the corresponding $s(t)$) and the non-sampled set $\underline{U}_{ns}$ (and the corresponding $U_{ns}(t)$) as

$$\underline{s} = \{k : k \in U,\, \underline{S}_k \geqslant 1\} \quad (s(t) = \{k : k \in U,\, S_k(t) \geqslant 1\}) \text{ and} \tag{2.6}$$

$$\underline{U}_{ns} = \{k : k \in U,\, \underline{S}_k = 0\} \quad (U_{ns}(t) = \{k : k \in U,\, S_k(t) = 0\}). \tag{2.7}$$

The difference between sample $\mathbf{S}(t)$ and sample set $s(t)$ is such: $s(t)$ is a subset of U and its units are determined by $\mathbf{S}(t)$, which is a N-dimensional vector of nonnegative integers.

Sampling can be with replacement (WR) and without replacement (WOR). In WOR sampling, units can be sampled only once ($\underline{S}_k = 0$ or $\underline{S}_k = 1$, $k = 1, \ldots, N$) and in WR sampling, more than once. We are interesting just in sampling without replacement in this research.

The distribution of $\underline{\mathbf{S}}$, denoted by $p(.)$, is called a sample design. The sampling design assigns a probability $\mathbf{P}(\underline{\mathbf{S}} = \mathbf{S}(t)) = p(\mathbf{S}(t))$ for every sample in time t. First and second order inclusion probabilities $\pi_k(t)$ and $\pi_{kl}(t)$ for sampling without replacement are defined as

$$\pi_k(t) = \mathbf{P}(\underline{S}_k = 1) = \sum_{\mathbf{S}(t):S_k=1} p(\mathbf{S}(t)), \tag{2.8}$$

$$\pi_{kl}(t) = \mathbf{P}(\underline{S}_k = 1, \underline{S}_l = 1) = \sum_{\mathbf{S}(t):S_k,S_l=1} p(\mathbf{S}(\mathbf{t})). \tag{2.9}$$

The design where first and second-order inclusion probabilities are strictly positive is design measurable. For every sampling design $\pi_{kk}(t) = \pi_k(t)$ and for WOR designs,

$$\begin{aligned} \pi_k(t) &= E(S_k), \quad \pi_{kl}(t) = E(S_k, S_l), \\ cov(S_k, S_l) &= \Delta_{kl}(t) = \pi_{kl}(t) - \pi_k(t)\pi_l(t). \end{aligned} \tag{2.10}$$

The sampling weights for WOR designs are defined as

$$w_k(t) = \begin{cases} \pi_k^{-1}, & \text{if } k \in s(t); \\ 0, & \text{if } k \in U_{ns}(t). \end{cases} \tag{2.11}$$

Depending on the sampling design the sample size

$$n(t) = \sum_{k \in U} S_k(t), \tag{2.12}$$

can be random or non-random. If $n(t)$ is non-random, the sampling design is fixed-size. The sample size and the sample set in domain $U^{(d)}$ are

$$n^{(d)}(t) = \sum_{k \in U^{(d)}} S_k(t), \quad \text{and} \quad s^{(d)}(t) = s(t) \cap U^{(d)}. \tag{2.13}$$

There are two types of domains:

1. Planned domains. (Singh, Gambino and Mantel (1994)) For planned domains the sample size $n^{(d)}(t)$ in domain sample is fixed in advance, so really these domains are strata with possible different allocations.
2. Unplanned domains. If the sample size $n^{(d)}(t)$ in domain sample is random, domains are unplanned. The disadvantage of unplanned domains is that, there might be domains with zero elements in the sample $\mathbf{S}(t)$.

In this research domains are unplanned. It is assumed that the number of the elements in each domain $U^{(d)}$, $d = 1, \dots, D$, is known, but the domains are not used in the sample design. This means that the sample set in each domain, $s^{(d)}(t)$, has a random size.

In this research two special cases of the general sample design are considered: simple random sampling without replacement (SRSWOR) and probability proportional to size sampling. SRSWOR is a fixed-size design under which the inclusion probabilities in time t are constants:

$$\pi_k(t) = \frac{n(t)}{N} = f(t), \quad \pi_{kl}(t) = f(t)\frac{n(t)-1}{N-1}. \tag{2.14}$$

Here ratio $n(t)/N$ is sampling fraction, denoted by $f(t)$.

In probability proportional to size sampling the inclusion probabilities satisfy $\pi_k(t) \propto x_k(t)$ for some $x(t)$ whose values are known for every unit in the population. The probability proportional to size sampling without replacement is denoted as πPS. We consider only fixed size πPS designs. In πPS designs first and second-order probabilities need to be strictly positive and $\Delta_{kl}(t) < 0$. This allows the construction of unbiased variance estimator for simple linear estimators.

In some cases sampling is done with stratification. It means, that population U is divided into H nonoverlapping stratum U_h. The number of elements in each stratum is N_h, $h = 1, \dots, H$. In each stratum the units to the sample are selected separately. That means, that in each stratum the different sample design can be used, but in practice for all stratum designs are the same. We will consider two cases of stratified sampling: stratified simple random sampling without replacement (SSRSWOR) and stratified probability proportional to size sampling without replacement (SπPS). In these cases for all stratum the same sample design (SRSWOR or πPS) is used.

Once the sample is selected, the values of variable $y_k(t)$ are recorded for the units $k \in s(t)$ in time t. This information is collected into the vector $Y_s(t)$. The unobserved values for $U_{ns}(t)$ are collected into the vector $Y_{ns}(t)$. Auxiliary information, domain indicator matrix and sampling weights $\mathbf{w}(t) = (w_1(t), w_2(t), \dots, w_N(t))$ are divided in a similar way to a sampled and non-sampled parts. The data matrix (2.2) can be written

$$\mathbf{A}(t) = \begin{bmatrix} \mathbf{A}_s(t) \\ \mathbf{A}_{ns}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_s(t) & \mathbf{Q}_s & \mathbf{w}_s(t) & Y_s(t) \\ \mathbf{X}_{ns}(t) & \mathbf{Q}_{ns} & 0 & Y_{ns}(t) \end{bmatrix}. \tag{2.15}$$

Here instead of $\mathbf{w}_{ns}(t)$ is written 0, because we are interested just in WOR design, thus $w_k(t)$ satisfy (2.11) equation. Once the information about study variable $y(t)$ is collected, the parameter of interest might be estimated.

2.1.3. Estimator, estimate and accuracy of estimator

An estimator is a rule or algorithm that defines how to estimate the parameter of interest (in our case: domain total). It is a random variable, whose value depends on the sample and auxiliary information. An estimate is the realization of an estimator. In general, an estimator and estimate are denoted as $\hat{\theta}(\underline{\mathbf{S}})$ and $\hat{\theta}(\mathbf{S}(\mathbf{t}))$, or briefly $\underline{\hat{\theta}}(t)$ and $\hat{\theta}(t)$. For parameter $TOT^{(d)}(t)$, the estimator and estimate are $\underline{\widehat{TOT}}^{(d)}(t)$ and $\widehat{TOT}^{(d)}(t)$.

The estimator is accurate if its bias and variance are small. The bias is a difference between the parameter expectation and the value:

$$BIAS(\underline{\hat{\theta}}(t)) = E(\underline{\hat{\theta}}(t)) - \theta(t). \tag{2.16}$$

If the $BIAS(\underline{\hat{\theta}}(t)) = 0$, the estimator is unbiased. The bias might come with respect to design or model. The symbols E_p, var_p denote, respectively, expected value and variance under the sample design. They are defined as

$$E_p(\underline{\hat{\theta}}(t)) = \Sigma_{\mathbf{S}(t)\in\mathcal{S}} p(\mathbf{S}(t))\hat{\theta}(t), \quad \text{and} \tag{2.17}$$

$$var_p(\underline{\hat{\theta}}(t)) = \Sigma_{\mathbf{S}(t)\in\mathcal{S}} p(\mathbf{S}(t))[\hat{\theta}(t) - E_p(\underline{\hat{\theta}}(t))]^2. \tag{2.18}$$

The symbols $E_{\mathcal{M}}$, $var_{\mathcal{M}}$ denote, respectively, an expected value and a variance under the model $\mathcal{M}$. Thus the estimator might be model-unbiased or design-unbiased.

2.1 definition. *An estimator $\hat{\underline{\theta}}(t)$ is said to be model-unbiased in time t if $E_{\mathcal{M}}(\hat{\underline{\theta}}(t) - \theta(t)) = 0$.*

2.2 definition. *An estimator $\hat{\underline{\theta}}(t)$ is said to be design-unbiased in time t if $E_p(\hat{\underline{\theta}}(t)) - \theta(t) = 0$.*

The other accuracy measures are:

- mean square error

$$\begin{aligned} MSE_.(\hat{\underline{\theta}}(t)) &= E_.(\hat{\underline{\theta}}(t) - \theta(t))^2 = \\ &= var_.(\hat{\underline{\theta}}(t)) + [BIAS_.(\hat{\underline{\theta}}(t))]^2; \end{aligned} \tag{2.19}$$

- standard error

$$SE_.(\hat{\underline{\theta}}(t)) = \sqrt{var_.(\hat{\underline{\theta}}(t))}; \tag{2.20}$$

- coefficient of variation

$$cv_.(\hat{\underline{\theta}}(t)) = \frac{\sqrt{var_.(\hat{\underline{\theta}}(t))}}{E_.(\hat{\underline{\theta}}(t))}. \tag{2.21}$$

Here subscript p or $\mathcal{M}$ should be used instead of . and the accuracy measures will measure with respect to the sampling design or to the model. Of course there are measures, which measure with respect to the sampling design and to the model. One of them is called anticipated mean-squared error.

2.3 definition. *The anticipated mean-squared error of an estimator $\hat{\underline{\theta}}(t)$ is defined by*

$$MSE_{\mathcal{M}p}(\hat{\underline{\theta}}(t)) = E_{\mathcal{M}}E_p(\hat{\underline{\theta}}(t) - \theta(t))^2. \tag{2.22}$$

The anticipated mean-squared error is very useful, when the goal is to construct an optimal estimation strategy.

2.4 definition. *A strategy is a pair $\{p(\mathbf{S}(t)), \hat{\underline{\theta}}(t)\}$ comprising a sample design and an estimator.*

Nedyalkova, Tille (2008) use such definition of optimal model-assisted strategy:

2.5 definition. *An optimal model-assisted strategy is one with a design-unbiased estimator that, subject to*

$$\sum_{k \in U} \pi_k = n, \quad 0 < \pi_k \leqslant 1, \tag{2.23}$$

minimizes the anticipated mean-squared error of that estimator.

2.1.4. Nonresponse

Nonresponse is present in almost all surveys and special estimation techniques are required to deal with the problem. Nonresponse means that the desired data are not obtained for the entire sample set $s(t)$.

There are two types of nonresponse: unit nonresponse and item nonresponse. Let us say we are interesting not in a study variable $y(t)$, but in the study variables $y_1(t),\ y_2(t), \dots,\ y_j(t)$. Then

1. The element k is an unit nonresponse element if the entire vector of $y(t)$-values, $\mathbf{y_k}(t) = (y_{1k}(t), y_{2k}(t), \dots, y_{jk}(t))$, in time t is missing.
2. The element k is an item nonresponse element if at least one, but not all j components of the vector $\mathbf{y_k}(t) = (y_{1k}(t),\ y_{2k}(t), \dots,\ y_{jk}(t))$ in time t are missing.

In this research we focus on one study variable $y(t)$, thus the item nonresponse is the same as the unit nonresponse and will be called just nonresponse.

We denote by $\underline{r}$ the response set (and its realization $r(t)$) and by $\underline{s}_{nr}$ nonresponse set (and the corresponding $s_{nr}(t)$):

$$\underline{r} = \{k : k \in \underline{s}, \text{ and } y_k(t) \text{ is recorded}\} \tag{2.24}$$

$$(r(t) = \{k : k \in s(t), \text{ and } y_k(t) \text{ is recorded}\}) \text{ and} \tag{2.25}$$

$$\underline{s}_{nr} = \{k : k \in \underline{s}, \text{ and } y_k(t) \text{ is not recorded}\} \tag{2.26}$$

$$(s_{nr}(t) = \{k : k \in s(t), \text{ and } y_k(t) \text{ is not recorded}\}). \tag{2.27}$$

Once the sample is selected, the values of variable $y_k(t)$ are recorded for the units $k \in r(t)$. This information is collected into the vector $Y_r(t)$. The selected, but not recorded values for $s_{nr}(t)$ are collected to the vector $Y_{nr}(t)$ and the unobserved values for $U_{ns}(t)$ are collected into the vector $Y_{ns}(t)$. Auxiliary information, domain indicator matrix and sampling weights are divided in a similar way to a recorded, sampled but not recorded and non-sampled parts. Thus the data matrix (2.2) can be written as:

$$\mathbf{A}(t) = \begin{bmatrix} \mathbf{A}_r(t) \\ \mathbf{A}_{nr}(t) \\ \mathbf{A}_{ns}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_r(t) & \mathbf{Q}_r & \mathbf{w}_r(t) & Y_r(t) \\ \mathbf{X}_{nr}(t) & \mathbf{Q}_{nr} & \mathbf{w}_{nr}(t) & Y_{nr}(t) \\ \mathbf{X}_{ns}(t) & \mathbf{Q}_{ns} & 0 & Y_{ns}(t) \end{bmatrix}. \tag{2.28}$$

2.2. Types of estimators

2.2.1. Horvitz-Thompson estimator

Horvitz-Thompson (HT) estimator was developed by Narain (1951) and Horvitz and Thompson (1952). Using this estimator the domain total in time t can be estimated by this formula:

$$\widehat{\underline{TOT}}^{(d)}_{HT}(t) = \sum_{k \in s^{(d)}(t)} w_k(t) y_k(t). \tag{2.29}$$

HT estimator is a design-based estimator. Design-based estimators use information about the sampling design by the means of sampling weights. HT estimator is design unbiased by definition:

$$\begin{aligned} E_p(\widehat{\underline{TOT}}^{(d)}_{HT}(t)) &= E_p\Big(\sum_{k \in s^{(d)}(t)} w_k(t) y_k(t) \Big) = \qquad (2.30) \\ &= E_p\Big(\sum_{k \in U^{(d)}} w_k(t) y_k(t) S_k(t) \Big) = \\ &= \sum_{k \in U^{(d)}} w_k(t) y_k(t) E_p(S_k(t)) = \\ &= \sum_{k \in U^{(d)}} \frac{y_k(t)}{\pi_k(t)} \pi_k(t) = \sum_{k \in U^{(d)}} y_k(t) = TOT^{(d)}. \end{aligned}$$

It's variance estimator is

$$\widehat{var}_p(\widehat{\underline{TOT}}^{(d)}_{HT}(t)) = \sum_{k\in s^{(d)}(t)} \frac{1-\pi_k(t)}{\pi_k^2(t)} y_k^2(t) + \tag{2.31}$$

$$+ \sum_{k\in s^{(d)}(t)} \sum_{\substack{l \in s^{(d)}(t) \\ l \neq k}} \frac{\pi_{kl}(t)-\pi_k(t)\pi_l(t)}{\pi_k(t)\pi_l(t)} \frac{y_k(t)y_l(t)}{\pi_{kl}(t)}.$$

2.2.2. GREG-type estimators

Following Lehtonen, Särndal and Veijanen (2003), the generalized regression type (GREG-type) estimator in time t, may be expressed under the general form

$$\widehat{\underline{TOT}}^{(d)}_{GREG}(t) = \sum_{k\in U^{(d)}} \hat{y}_k(t) + \sum_{k\in s^{(d)}(t)} w_k(t)(y_k(t) - \hat{y}_k(t)). \tag{2.32}$$

where $\hat{y}_k(t)$ denotes the prediction of $y_k(t)$ under the assumed super population model (see section 2.6). The predictions, $\{\hat{y}_k(t); k \in U\}$ differ from one model specification to another, depending on the functional form and from the choice of the auxiliary variables.

The GREG-type estimator is a design-based model-assisted estimator, which is nearly design unbiased irrespective of the model choice. Here a statistical model is used as an assisting tool to incorporate auxiliary information into the estimation procedure.

If a working super population model is

$$\underline{y}_k(t) = \underline{\mathbf{x}}'_k(t)\beta^{(d)}(t) + \underline{\varepsilon}_k(t). \tag{2.33}$$

The predictions $\hat{y}_k(t)$ are then obtained by

$$\hat{y}_k(t) = \mathbf{x}'_k(t)\hat{\beta}^{(d)}(t) \tag{2.34}$$

being

$$\hat{\beta}^{(d)}(t) = \Big(\sum_{k\in s^{(d)}(t)} \mathbf{x}_k(t)\mathbf{x}'_k(t)w_k(t)\Big)^{-1} \sum_{k\in s^{(d)}(t)} \mathbf{x}_k(t)y_k(t)w_k(t). \tag{2.35}$$

In this case GREG-type estimator might be expressed as:

$$\begin{aligned}
\widehat{\underline{TOT}}^{(d)}_{GREG}(t) &= \sum_{k\in s^{(d)}(t)} w_k(t)y_k(t) + \sum_{k\in U^{(d)}} \mathbf{x}'_k(t)\hat{\beta}^{(d)}(t) \\
&\quad - \sum_{k\in s^{(d)}(t)} w_k(t)\mathbf{x}'_k(t)\hat{\beta}^{(d)}(t) = \sum_{k\in s^{(d)}(t)} w_k(t)y_k(t) + \\
&\quad + \Big(\sum_{k\in U^{(d)}} \mathbf{x}_k(t) - \sum_{k\in s^{(d)}(t)} \mathbf{x}_k(t)w_k(t)\Big)'\hat{\beta}^{(d)}(t) = \\
&= \sum_{k\in s^{(d)}(t)} \Big[1 + \Big(\sum_{k\in U^{(d)}} \mathbf{x}_k(t) - \sum_{k\in s^{(d)}(t)} \mathbf{x}_k(t)w_k(t)\Big)' \times \\
&\times \Big(\sum_{k\in s^{(d)}(t)} \mathbf{x}_k(t)\mathbf{x}'_k(t)/\sigma^2_k(t)\pi_k(t)\Big)^{-1} \mathbf{x}_k(t)/\sigma^2_k(t)\Big] \times \\
&\times w_k(t)y_k(t) = \sum_{k\in s^{(d)}(t)} a_k(t)y_k(t).
\end{aligned}$$

Such estimator is the same as a calibrated estimator (Deville and Särndal (1992)) if function

$$L = \sum_{k\in s^{(d)}(t)} \frac{(a_k(t) - w_k(t))^2}{w_k(t)} \tag{2.36}$$

is used for minimizing distance. Here $a_k(t)$ is called calibrated weight.

If $\mathbf{x}_k = (1, x_k)'$ in (2.33) then the GREG-type estimator is equal to regression estimator (Deville and Särndal (1992)), and if $\mathbf{x}_k = x_k$, the GREG-type estimator is equal to ratio estimator (Deville and Särndal (1992)).

As noted by Rao (2003) the GREG-type estimator under the superpopulation model (2.33) is approximately design unbiased as the overall sample size increases, even if the domain sample size $n^{(d)}$ is small. Moreover, the sum of the $\widehat{TOT}^{(d)}_{GREG}(t)$ estimates over all the domains of a partitions is benchmarked to the usual GREG estimate of the total

$$\begin{aligned}
\sum_{d=1}^{D} \widehat{TOT}^{(d)}_{GREG}(t) &= \sum_{k\in s(t)} \Big[1 + \Big(\sum_{k\in U} \mathbf{x}_k(t) - \sum_{k\in s(t)} \mathbf{x}_k(t)w_k(t)\Big)' \times \\
&\times \Big(\sum_{k\in s(t)} \mathbf{x}_k(t)\mathbf{x}'_k(t)/\sigma^2_k(t)\pi_k(t)\Big)^{-1} \mathbf{x}_k(t)/\sigma^2_k(t)\Big] w_k(t)y_k(t).
\end{aligned}$$

The estimator of the variance of the GREG estimator in time t can be expressed as

$$\widehat{var}(\widehat{TOT}^{(d)}_{GREG}(t)) = \sum_{k\in s^{(d)}(t)} \sum_{l\in s^{(d)}(t)} \Big(1 - \frac{\pi_k(t)\pi_l(t)}{\pi_{kl}(t)}\Big)\times$$

$$\times \frac{y_k(t) - \hat{y}_k(t)}{\pi_k(t)} \frac{y_l(t) - \hat{y}_l(t)}{\pi_l(t)}. \tag{2.37}$$

2.2.3. Model-based estimators

If sample sizes are too small to apply direct survey estimators and additional information is available, model-dependent or model-based estimation procedures might be used to produce sufficiently reliable statistics. In these procedures a model is applied to borrow information from other related data sets to improve the precision of the estimates.

The general form of a model-based estimator for domain is equal to

$$\widehat{TOT}^{(d)}_{MB}(t) = \sum_{k\in U^{(d)}\setminus s^{(d)}} \hat{y}_k(t) + \sum_{k\in s^{(d)}} y_k(t). \tag{2.38}$$

Therefore, in (2.38), the observed sample speaks for itself and the rest is predicted according to the fitted model.

For the model-based estimators the sampling weights are used in the calculation of superpopulation model's coefficients. If the sampling weights are equal $w_k(t) = w(k)$ for all $k \in s$, then model-based estimators are equal to the model-dependent estimators (2.38). These estimators do not incorporate the sampling weights in the calculation of superpopulation model's coefficients.

The name of estimator (2.38) varies under the different superpopulation model (Rao (2003)). For example, let the superpopulation model is

$$\underline{Y}(t) = \underline{\mathbf{X}}(t)\beta(t) + \underline{\mathbf{Z}}(t)\mathbf{v}(t) + \underline{\varepsilon}(t),$$

where $\underline{\mathbf{X}}(t)$ and $\underline{\mathbf{Z}}(t)$ are known $n \times J$ and $n \times p$ matrices, $\mathbf{v}(t)$ and $\underline{\varepsilon}(t)$ are independently distributed with mean 0 and covariances matrices $\mathbf{G}(t)$ and $\mathbf{R}(t)$ depending on some variances parameters $\delta(t) = \{\delta_1(t), \delta_2(t), \ldots, \delta_q(t)\}$. If a sampling weights are used in the calculation of superpopulation model's coefficients, then the estimator (2.38) is called a pseudo empirical best linear unbiased predictor (pseudo-EBLUP estimator) (Rao (2003)).

Compared to the traditional design-based survey estimators, model-based estimators have much smaller variances. The price that is paid for this variance reduction is that these model-based estimators are more or less design-biased. The size of the bias depends on the correctness of the model.

2.2.4. Benchmarking

Direct survey estimates are often adequate at an aggregate (or large area) level in terms of precision. It is, therefore, sometimes desirable to modify the individual small area estimators so that a properly weighted sum of these estimators equals to the model-free, direct estimator at the aggregate level. The modified estimators will be somewhat less efficient than the original, optimal estimators, but they avoid possible aggregation bias by ensuring consistency with the direct estimator.

One simple way to achieve consistency is to make a ratio adjustment, for example, the model-based estimator $\widehat{TOT}^{(d)}_{MB}(t)$ of total $TOT^{(d)}(t)$ is modified to

$$\widehat{TOT}^{(d)}_{mod}(t) = \frac{\widehat{TOT}^{(d)}_{MB}(t)}{\sum_d \widehat{TOT}^{(d)}_{MB}(t)} \widehat{TOT}_{dir}(t), \qquad (2.39)$$

where $\widehat{TOT}_{dir}(t)$ is a direct estimator of the aggregate population total

$$TOT(t) = \sum_d TOT^{(d)}(t).$$

2.3. Classification of estimators

There are several ways to classify estimators. One of the most popular is estimator classification by the use of the super population model (see section 2.3.1). The other way to classify estimators is by the what auxiliary information is used (see section 2.3.2). The last way of classification, which is mention in this research is the classification by the estimator's form (see section 2.3.3).

2.3.1. Design-based, model-depending and model-based estimators

The purpose of survey sampling is to obtain statistical information about a finite population by selecting a probability sample from this population, measuring the required information about the units in this sample and estimating finite population parameters such as means, totals and ratios. The statistical inference in this setting can be design-based, model-assisted, model-depending or model-based.

In the design-based and model-assisted approach, the probability structure for inferences comes from the randomization distribution, from the probabilities with which different samples are potentially drawn (although one and only one is realized in a survey). The statistical properties (mean, variance and so on) of an estimate are evaluated by averaging over all possible samples under the given sampling design. Here statistical modeling plays a minor role.

In the model-based and model-dependent context, the probability structure of the sampling design plays a less pronounced role, since the inference is based on the probability structure of an assumed statistical model.

Design-based and model-assisted estimators refer to a class of estimators that expand or weight the observations in the sample with the so-called sampling weights. Sampling weights are derived from the sampling design and available auxiliary information about the target population. A well known design-based estimator is Horvitz-Thompson estimator (see section 2.2.1). GREG-type estimators (see section 2.2.2) are design-based model-assisted estimators. These estimators are derived from model that specifies the relationship between the values of a certain target parameter and a set of auxiliary variables for which the totals in the finite target population are known. After these estimators are derived, they are judged by their design-based properties, such as design expectation and design variance.

If the underlying model of the GREG-type estimator explains the variation of the target parameter in the finite population reasonably well, then this might result in a reduction of the design variance of the Horvitz-Thompson estimator. If the model is misspecified, then this might increase of the design variance but the property that the GREG-type estimator is approximately design unbiased remains. From this point of view, the GREG-type estimator is robust against model-misspecification.

Model-dependent and model-based estimators (see section 2.2.3) refer to the class of estimators that use models to estimate certain target parameter. The model-dependent estimators does not use design information at all, thus for estimating these estimators it is not needed to have probability sample. As for the model-based estimators sampling weights are used to estimate model's coefficients. Compared with design-based or model-assisted estimators, model-based and model-dependent estimators have much smaller variances, but they are more or less design-biased. The size of the bias depends on the correctness of the model and in many cases it does not decrease when sample size increases.

2.3.2. Direct and indirect estimators

Estimators for domains are frequently classified as either direct or indirect. In the terminology of Schaible (1992) and Federal Committee on Statistical Methodology (1993), an estimator for a domain is called direct only if it uses values of the variable of interest over the domain and for the time period in question. Otherwise, it is indirect.

A convenient direct estimator is Horvitz-Thompson (HT) estimator (2.29). The other estimators, used in this research are direct or indirect depending on the superpopulation model used in estimation stage.

2.3.3. Linear and nonlinear estimators

A linear estimator can be described as

$$\widehat{TOT}^{(d)}(t) = \sum_{k \in s^{(d)}(t)} a_k(t) y_k(t) = \sum_{k \in U} a_k(t) y_k(t) I_k^{(d)}, \qquad (2.40)$$

where the a_k, $u \in s^{(d)}(t)$ are weights that can depend on the sample and $I_k^{(d)}$ is equal to 1 if $k \in s^{(d)}(t)$ and equal 0 otherwise.

All design-based and model-assisted estimators are linear estimators.

The estimators, which can't be expressed by (2.40) equation are called nonlinear estimators.

2.4. Nonresponse adjustment using weighting methods

Weighting adjustment is a popular method for handling unit nonresponse in sample surveys. Groves, Dillman, Eltinge and Little (2002) and Särndal, Lundstrom (2005) provided comprehensive overviews of nonresponse weighting adjustment (NWA) methods in survey sampling. The primary objective of a weight adjustment procedure is to reduce the nonresponse bias, which is introduced when respondents and nonrespondents are different with respect to the survey variables.

Using weighting method the original inclusion probabilities $\pi_k(t)$ are deflated by the response probabilities $\varkappa_k(t)$ and new sampling weights $w_k(t) = (\pi_k(t)\varkappa_k(t))^{-1}$, $k \in r(t)$, are obtained.

2.4.1. Estimation of nonresponse probability

The original response probability is never known in practice, so there are several methods to estimate it.

A very popular method for estimation of nonresponse probability in practice is called a weighting-class. It consists of first dividing the the response sample set $r(t)$ and sample set $s(t)$ into G mutually exclusive weighting classes $r_g(t)$ and $s_g(t)$, $g = 1, \ldots, G$, and adjusting the design weights of respondents by the inverse of the response rate within each class. These classes are formed on the basis of auxiliary information recorded for all units in the sample (see, Little (1986) and Eltinge and Yansaneh (1997)).

The estimate of the response probability $\varkappa_k(t)$ for the unit is the same in the same class:

$$\hat{\varkappa}_k(t) = \frac{\sum_{j \in r_g(t)} w_j(t)}{\sum_{j \in s_g(t)} w_j(t)}, \quad k \in r(t). \tag{2.41}$$

Another method for estimating the response probability is to apply a logistic regression model (Ekholm and Laaksonen (1991)):

$$\hat{\varkappa}_k(t) = \frac{exp\{\hat{B}(t)\mathbf{z}_k(t)\}}{1 + exp\{\hat{B}(t)\mathbf{z}_k(t)\}}, \quad k \in r(t). \tag{2.42}$$

Here $\hat{B}(t)$ is the maximum likelihood estimator of the coefficients of the logistic regression model based on the data $\{(I_{r,k}(t),\, \mathbf{z}_k(t)), k \in \mathbf{s}\}$ where $I_{r,k}(t) = 1$, if $k \in r(t)$, and $I_{r,k}(t) = 0$ otherwise. The auxiliary information is notated as $\mathbf{z}_k(t)$, because it might be different from the auxiliary information which will be used in estimation stage. Ekholm and Laaksonen (1991) suggested response probabilities are model-based estimates ($0 < \hat{\varkappa}_k(t) \leq 1$).

2.4.2. Estimation of domain total when weighting methods are used

When weighting methods for nonresponse adjustment are applied in the estimation of the domain total, the correction of estimators should be made by replacing sampling weights $w_k(t)$ with $\hat{w}_k(t) = (\pi_k(t)\hat{\varkappa}_k(t))^{-1}$, $k \in r(t)$, not only in estimators equations, but also in calculation of the coefficients of the model. Thus the HT estimator is equal to

$$\widehat{\underline{TOT}}^{(d)}_{HT,weight}(t) = \sum_{k \in r^{(d)}(t)} \hat{w}_k(t) y_k(t), \tag{2.43}$$

GREG-type estimator is equal to

$$\widehat{\underline{TOT}}^{(d)}_{GREG,weight}(t) = \sum_{k \in U^{(d)}} \hat{\hat{y}}_k(t) + \sum_{k \in r^{(d)}(t)} \hat{w}_k(t)(y_k(t) - \hat{\hat{y}}_k(t)) \tag{2.44}$$

and model-based estimator to

$$\widehat{TOT}^{(d)}_{MB,weight}(t) = \sum_{k \in U^{(d)} \setminus r^{(d)}} \hat{\hat{y}}_k(t) + \sum_{k \in r^{(d)}} y_k(t). \tag{2.45}$$

Here $r^{(d)}(t) = r(t) \cup s^{(d)}(t)$ and $\hat{\hat{y}}_k(t)$ are predicted values of $y_k(t)$ under the assumed super population model (see section 2.6). The difference between $\hat{\hat{y}}_k(t)$ and $\hat{y}_k(t)$ is that the model's coefficients for $\hat{\hat{y}}_k(t)$ are estimated using just response set $r(t)$, but not the sample set $s(t)$ as it is done for $\hat{y}_k(t)$.

If the response probability is estimated using weighting classes, the HT estimator, which use new sampling weights is unconditionally unbiased, but conditional bias can arise when a difference between the distribution of a weighted population and the sample level based weighted population exists. Oh and Scheuren (1983) showed that conditional bias cannot be directly derived. However, an average can be obtained (Kalton and Maligalig (1991)).

If the response probability is estimated using Ekholm and Laaksonen (1991) suggested method and if nonresponse is believed to be ignorable in each adjustment cell and the applied model is correct in explaining the true response distribution, then the estimator 2.43 is asymptotically unbiased.

2.5. Nonresponse adjustment using imputation methods

Imputation methods can be classified as a single imputation (when one value is imputed instead of missing one) or multiple imputation. Multiple imputation produces several imputed datasets and instead of the missing value a mean of imputed datasets is used.

There are many types of imputation methods, which can be divided into three main groups:

1. Logical (deductive) imputation, when the imputed value is calculated using logical assumptions.
2. Real donor imputation, where the imputed observation value is borrowed from another respondent.
3. Model-based imputation, where the imputed value is calculated using the model with the coefficients estimated from the response sample $r(t)$.

In this research logical imputation was not used. In the section 2.5.1 the main imputation methods which are used in this research are described. How the estimators change when imputation methods are used for nonresponse adjustment are described in section 2.5.2.

2.5.1. Imputation methods

A once-common method of imputation using donors is a hot-deck imputation in which each missing value is replaced with an observed response from a "similar" unit. The term "hot deck" dates back to the storage of data on punched cards, and indicates that the information donors come from the same dataset as the recipients. The stack of cards was "hot" because it was currently being processed. Cold-deck imputation, by contrast, selects donors from another dataset (e.g. previous surveys).

Hot deck imputation involves replacing missing values of one or more variables for a non-respondent (called the recipient) with observed values from a respondent (the donor) that is similar to the non-respondent with respect to characteristics observed by both cases. Here we review two different forms of the hot deck imputation.

In the first form, the donor is selected randomly from a set of potential donors, which can be called the adjustment cells. Such method is called random donor method. The adjustment cells are based on auxiliary variables which are known and for donors and for recipients. The choice of auxiliary variables for creating adjustment cells often relies on subjective knowledge of which variables are associated with the item being imputed, and predictive of nonresponse. Imputation is then carried out by randomly picking a donor for each non-respondent within each cell.

In the second form, a single donor is identified for each recipient by using some metric. This method is called a nearest neighbor. For the nearest neighbor imputation, a missing value $y_k(t)$ is imputed by choosing that value $y_l(t)$ which corresponds to the value $\mathbf{x}_l(t)$ closest to $\mathbf{x}_k(t)$. The closest value is determined by the distance between any two response values:

$$d_{kl}(t) = \sqrt{\sum_{j=1}^{J}(x_{kj}(t) - x_{lj}(t))^2}, \quad k \in s(t)\backslash r(t), \quad l \in r(t). \tag{2.46}$$

This procedure can be done if the continuous variables are use to identify the distant.

The other group of imputation methods are model-based imputation. A most common model-based method is regression model. Here an imputation model predicts a missing value using a function of some auxiliary variables. The auxiliary variables can be from the same survey, or from the other sources. The regression coefficients can be determined using response set from the current survey or from historic survey data.

2.5.2. Estimation of domain total when imputation is used

Let us denote a new variable $y^*(t)$ which values $y_k^*(t)$ are equal to $y_k(t)$, if $k \in r(t)$, or $y_k^{imp}(t)$, if $k \in s(t)\backslash r(t)$. Here y_k^{imp} can be a single imputed value, if single imputation is used, or the mean of imputed datasets, if multiple imputation is used.

Thus the HT estimator is equal to

$$\widehat{\underline{TOT}}^{(d)}_{HT,imp}(t) = \sum_{k\in s^{(d)}(t)} w_k(t) y_k^*(t), \tag{2.47}$$

GREG-type estimator is equal to

$$\widehat{\underline{TOT}}^{(d)}_{GREG,imp}(t) = \sum_{k\in U^{(d)}} \hat{y}_k^*(t) + \sum_{k\in s^{(d)}(t)} w_k(t)(y_k^*(t) - \hat{y}_k^*(t)) \tag{2.48}$$

and model-based estimator to

$$\widehat{TOT}^{(d)}_{MB,imp}(t) = \sum_{k\in U^{(d)}\backslash s^{(d)}} \hat{y}_k^*(t) + \sum_{k\in s^{(d)}} y_k^*(t). \tag{2.49}$$

2.6. Models in small area estimation

If no other data sources are available, statisticians can only resort to model-based methods which involve making assumptions about how data for a small area relate to other data. These methods are often described as “borrowing strength” since they borrow information from elsewhere in the sample survey to augment the number of units that contribute to the estimate for a given small area. The borrowing can be from other time periods, from sample units outside the given small area, or from other variables measured on the same sample unit.

There are two types of models used in small area estimation: area level model and unit level model.

2.6.1. Area level model

For the area level model only area-specific auxiliary data $\mathbf{x}^{(d)}(t) = (x_1^{(d)}(t), ..., x_J^{(d)}(t))'$ in time t are available for the sampled areas $d = 1, ..., D$ as well as the nonsampled areas and the parameters of interest, $TOT^{(d)}(t)$, are assumed to be related to $\mathbf{x}^{(d)}(t)$ through a linear model with random area effects:

$$TOT^{(d)}(t) = \mathbf{x}^{(d)}(t)'\beta(t) + \mathbf{v}^{(d)}(t), \quad d = 1, ..., D, \tag{2.50}$$

where $\beta(t)$ is the J-vector of regression parameters and the $\mathbf{v}^{(d)}(t)$’s are independent and identically distributed (IID) random variables with

$$E\Big(\mathbf{v}^{(d)}(t)\Big) = 0, \quad var\Big(\mathbf{v}^{(d)}(t)\Big) = \sigma_{\mathbf{v}}^2(t). \tag{2.51}$$

In addition, normality of the random effects $\mathbf{v}^{(d)}(t)$ is often assumed.

It is also possible to partition the areas into groups and assume separate models of the form (2.50) across groups.

It is assumed that direct estimators $\widehat{TOT}^{(d)}(t)$ of $TOT^{(d)}(t)$ are available whenever the area sample size $n(t) \geq 1$. It is also customary to assume that

$$\widehat{TOT}^{(d)}(t) = TOT^{(d)}(t) + e^{(d)}(t), \tag{2.52}$$

here the sampling errors $e^{(d)}(t)$ are independent $\mathcal{N}(0, \psi^{(d)}(t))$ with known $\psi^{(d)}(t)$. Combining this sampling model (2.52) with the “linking” model (2.50), the well-known area level linear mixed model of Fay and Herriot (1979) is got:

$$\widehat{TOT}^{(d)}(t) \,=\, \mathbf{x}^{(d)}(t)'\beta(t) + \mathbf{v}^{(d)}(t) + e^{(d)}(t). \tag{2.53}$$

Note that (2.53) involves both design-based random variables $e^{(d)}(t)$ and model-based random variables $\mathbf{v}^{(d)}(t)$. In practice, sampling variances $\psi^{(d)}(t)$ are seldom known, but smoothing of estimated variances $\hat{\psi}^{(d)}(t)$ is often done to get stable estimates $\psi^{*,(d)}(t)$ which are then treated as the true $\psi^{(d)}(t)$.

The basic area level model has been extended to handle correlated sampling errors, spatial dependence of random small area effects, vectors of parameters $\mathbf{TOT}^{(d)}$ (multivariate case), time series and cross-sectional data and others (see Ghosh and Rao (1994)).

In this research the auxiliary information is available not only at the area level, but also at the unit level. Also the sample designs are used so, that there is no nonsampled small areas, thus the area level model won't be used in the simulation part.

2.6.2. The basic unit level model

A basic unit level population model assumes that the unit $y(k)$-values, y_k are related to auxiliary variables $\mathbf{x}_k(t)$ through an one-way nested error regression model

$$y_k(t) \,=\, \mathbf{x}_k(t)'\beta^{(d)}(t) + \mathbf{v}^{(d)}(t) + e_k(t), \tag{2.54}$$

where $k \,=\, 1, ..., N$ and $d \,=\, 1, ..., D$. Here $\mathbf{v}^{(d)}(t) \overset{IID}{\sim} \mathcal{N}(0, \sigma_v^2(t))$ are independent of $e_k(t) \overset{IID}{\sim} \mathcal{N}(0, \sigma_e^2(t))$. The parameters of interest here are the small area totals $TOT^{(d)}(t)$.

It is possible to write model (2.54) in matrix form as

$$Y(t) \,=\, \mathbf{X}(t)\beta(t) + \mathbf{V}(t)\mathbf{1}^{(d)} + E(t), \tag{2.55}$$

where $\mathbf{X}(t)$ is $N \times J$, $Y(t)$, $E(t)$ are $N \times 1$ and $\mathbf{1}^{(d)} \,=\, (1, ..., 1)'$.

2.6.3. Generalized Unit level model

In this research a more general model than (2.54) is considered, namely a panel data model. This type of the model was considered, because, the study variable $y_k(t)$ and auxiliary variables $\mathbf{x}_k(t)$ are time series, thus in time $T+1$, when the estimators of parameters of interest are needed, there is a huge set of historical i.e. prior to the sample selection, auxiliary information

$$AI := (\mathbf{x}_k(t), y_k(t),\ t \in \mathcal{T}_k \subset \{1, 2, \ldots, T\},\ k \in U), \tag{2.56}$$

which might be used to improve estimation strategy. One of the way to use such kind of information is to use panel data model.

Below a general panel data model with random effects is given:

$$\begin{aligned} \underline{y}_k(t) \ &= \beta_{0,g(k)}(t) + v_{0,k}(t) + \sum_{j=1}^{J} [\beta_{j,g(k)}(t) + v_{j,k}(t)]\underline{x}_{j,k}(t) + \\ &+ \sum_{i=1}^{m} \alpha_{i,g(k)}\mu_i(t) + \varepsilon_k(t), \quad k \in U. \end{aligned} \tag{2.57}$$

Here $\underline{x}_{j,k}(t), j = 1, 2, ..., J$, 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 $v_{0,k}(t)$, $v_{1,k}(t)$,..., $v_{J,k}(t)$, $(v_{j,k}(t) \sim IID(0, \lambda^2_{0,g(k)}(t))$, $g(k) = 1(k), ..., G(k)$, $j = 0, ..., J$.

The model error is denoted as $\varepsilon_k(t)$ $(E_{\mathcal{M}}(\varepsilon_k(t)) = 0$, $var_{\mathcal{M}}(\varepsilon_k(t)) = \nu_k^2\sigma^2$, $\forall k \in U$ and $cov_{\mathcal{M}}(\varepsilon_k(t), \varepsilon_l(t)) = 0$ when $k \neq l$).

It should be noticed that model error $\varepsilon_k(t)$ and the random-effects model coefficients $v_{0,k}(t)$, $v_{1,k}(t)$,..., $v_{J,k}(t)$ are conditionally independent if values of $\underline{x}_{j,k}(t), j = 1, 2, ..., J$, 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.

2.6.4. Special cases of generalized Unit level model

In this section it is shown how from the generalized unit level model can be obtained some other well known models.

1. Example. Let $\beta_{0,g(k)}(t) = \beta_0(t)$, $v_{0,k}(t) = 0$, $\beta_{j,g(k)}(t) = \beta_j(t)$, $v_{j,k}(t) = 0$, $j = 1, ..., J$ and t is equal to one moment (let this moment is notated as W). Then the generalized unit level model has

such form

$$\underline{y}_k(W) = \beta_0(W) + \sum_{j=1}^{J} \beta_j(W)\underline{x}_{j,k}(W) + \varepsilon_k(W), \quad k \in U. \quad (2.58)$$

This model is known as a common model (Lehtonen, Särndal and Veijanen (2003)), because it has the same model coefficients for all domains.

2. Example. Let $\beta_{0,g(k)}(t) = \beta_0^{(d)}(t)$, $v_{0,k}(t) = 0$, $\beta_{j,g(k)}(t) = \beta_j(t)$, $v_{j,k}(t) = 0$, $j = 1, ..., J$ and t is equal to one moment (let this moment is notated as W). Then the generalized unit level model has such form

$$\underline{y}_k(W) = \beta_0^{(d)}(W) + \sum_{j=1}^{J} \beta_j(W)\underline{x}_{j,k}(W) + \varepsilon_k(W), \quad k \in U. \quad (2.59)$$

This model is known as a model with domain-intercept (Lehtonen, Särndal and Veijanen (2003)), because it has the same slopes but separate intercepts for all domains.

3. Example. Let $\beta_{0,g(k)}(t) = \beta_{0,g(k)}$, $v_{0,k}(t) = 0$, $\beta_{j,g(k)}(t) = \beta_{j,g(k)}$ $v_{j,k}(t) = 0$, $j = 1, ..., J$. Then the generalized unit level model has such form

$$\underline{y}_k(t) = \beta_{0,g(k)} + \sum_{j=1}^{J} \beta_{j,g(k)}\underline{x}_{j,k}(t) + \varepsilon_k(t), \quad k \in U. \quad (2.60)$$

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 to find the model coefficients just using data from the past. The current data might be use just for prediction.

4. Example. Let $\beta_{0,g(k)}(t) = \beta_0(t)$, $v_{0,k}(t) = v_{0,g(k)}(t)$, $\beta_{j,g(k)}(t) = \beta_j(t)$, $v_{j,k}(t) = 0$, $j = 1, ..., J$ and t is equal to one moment (let this moment is notated as W). Then the generalized unit level model has

such form

$$\begin{aligned}\underline{y}_k(W) =&\beta_0(W) + v_{0,g(k)}(W) + \\ &+ \sum_{j=1}^{J} \beta_j(W)\underline{x}_{j,k}(W) + \varepsilon_k(W), \quad k \in U. \end{aligned} \tag{2.61}$$

This model is known as mixed model with random-intercept (Lehtonen, Särndal and Veijanen (2003)), because it has the same fixed parameters for all domains and the random effect is defined at the group level, which in separate case might be equal to domain.

5. Example. Let $\beta_{0,g(k)}(t) = \beta_{0,g(k)}$, $v_{0,k}(t) = v_{0,g(k)}$, $\beta_{j,g(k)}(t) = \beta_{j,g(k)}$, $v_{j,k}(t) = 0$, $j = 1, ..., J$. Then the generalized unit level model has such form

$$\underline{y}_k(t) = \beta_{0,g(k)} + v_{0,g(k)} + \sum_{j=1}^{J} \beta_{j,g(k)}\underline{x}_{j,k}(t) + \varepsilon_k(t), \quad k \in U. \tag{2.62}$$

This model is mixed panel data model, which coefficients do not depend on time. Thus, the model coefficients might be calculated from the previous information and auxiliary variables from the current sample might be used just for prediction.

6. Example. Let $\beta_{0,g(k)}(t) = \beta_{0,g(k)}$, $v_{0,k}(t) = v_{0,g(k)}$, $\beta_{j,g(k)}(t) = \beta_{j,g(k)}$, $v_{j,k}(t) = 0$, $j = 1, ..., J$ and

$$\sum_{i=1}^{m} \alpha_{i,g(k)}\mu_i(t) = a_{0,g(k)}t + \mathbf{a}'_{g(k)}\alpha(t),$$

where $\mathbf{a}_{g(k)} \in \mathbf{R}^3$. Then the generalized unit level model has such form

$$\begin{aligned}\underline{y}_k(t) =&\beta_{0,g(k)} + v_{0,g(k)} + a_{0,g(k)}t + \mathbf{a}'_{g(k)}\alpha(t) + \\ &+ \sum_{j=1}^{J} \beta_{j,g(k)}\underline{x}_{j,k}(t) + \varepsilon_k(t), \quad k \in U. \end{aligned} \tag{2.63}$$

This is a mixed panel data model with a linear trend and a seasonal components.

2.7. The summary of the second chapter

1. All definitions and notations used in this research are defined in the second chapter.
2. The formulas of Horvitz-Thompson (HT), generalized regression (GREG) and model-based (MB) estimators are described for small area estimation case. Also all these formulas are adapted in respect of different nonresponse adjustment methods.
3. Different methods of estimators classification (design-based, model-based or model-dependent, direct or indirect, linear or nonlinear) are described and for each estimator (Horvitz-Thompson, GREG-type and model-based) the place in each classification is showed.
4. Two nonresponse adjustment types (nonresponse adjustment using weighting methods or imputation methods) are described.
5. Two types of models used in small area estimation are described. The first is called area level model and the second – unit level model. Unit level model is described in more detailed. It is showed, that it is possible to write it in a general form, from which main models used in small area estimation can be obtained.

3

Optimal estimation strategy

The small area problem is usually considered to be treated via estimation. 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.

As noted by Singh, Gambino and Mantel (1994), there is a need to develop an overall strategy that deals with small area problems, involving both planning sample design and estimation aspects.

In this chapter some other not so common used sample designs are considered. A definition of balanced sample is introduced in section 3.1. Some examples and methods how to select a balanced sample are also presented.

The other sample design is discussed in section 3.2. It is called model-based design, because for selecting the sample the variance of the prediction errors is used.

3.1. Balanced samples

Consider a sample $\mathbf{S}(t)$ of size $n(t)$ that is a subset of a finite population U of size N. A sample is said to be balanced if, for a vector of auxiliary variable

$\mathbf{z}_k = (z_{1,k}, ..., z_{p,k}, ..., z_{P,k})'$,

$$\frac{1}{n}\sum_{k\in s(t)} \mathbf{z}_k(t) = \frac{1}{N}\sum_{k\in U} \mathbf{z}_k(t), \qquad (3.1)$$

which means that the sample means of the z-variables match their population means. Here the auxiliary information is notated as $\mathbf{z}_k(t)$, because it might be different from the auxiliary information which will be used in estimation stage and which is notated as $\mathbf{x}_k = (x_{1,k}, ..., x_{j,k}, ..., x_{J,k})'$.

Brewer (1999) drew a distinction between a balanced selection of samples and a random selection of samples. However, a balanced sample may be selected randomly. If a random sample $\mathbf{S}(t)$ is selected randomly, then each unit of the population has an inclusion probability $\pi_k(k)$ of being selected. In this case, a random sample must satisfy the following balancing equation:

$$\sum_{k\in s(t)} \frac{\mathbf{z}_k(t)}{\pi_k(t)} = \sum_{k\in U} \mathbf{z}_k(t). \qquad (3.2)$$

In other words, in a balanced sample, the total of the z-variables are estimated without error. Really equation (3.1) is a special case of equation (3.2) (when $\pi_k(t) = n(t)/N$ or when the sample is not selected randomly), but several authors like Cumberland and Royall (1981) and Kott (1986) would call a sample that satisfies equation (3.2) a "π-balanced sample", and a sample that satisfies equation (3.1) as "mean-balanced sample".

Deville and Tille (2004) defined such definition of a balanced sampling design:

3.1 definition. *A sampling design $p(\cdot)$ is said to be balanced on auxiliary variables $z_1(t), ..., z_P(t)$ if the Horvitz-Thompson estimator satisfies Equation (3.2).*

A balanced sampling can be viewed as a kind of calibration (GREG-type estimators belongs to calibrated estimators class (see section 2.2.2)) that is directly integrated into the sampling design. The main problem is that the balancing equations (3.2) can rarely be exactly satisfied.

Still the advantages of balanced sampling are as follows:

1. Balanced sampling increases the accuracy of the Horvitz-Thompson estimator. The variance of the Horvitz-Thompson estimator only depends on the residuals of the regression of the interest variable by the balancing variables.

2. Balanced sampling protects against large sampling errors. The most unfavorable samples have a null probability of being selected when balanced sampling design is used.

3. If the study variable is well explained by the auxiliary information, in model-based inference, balanced sampling protects against a misspecification of the model. A recent discussion of this important question is given in Nedyalkova, Tille (2008).

4. If an indicator variable of the domain is added in the list of auxiliary variables, then the size of the domain is fixed in the sample and this protect domains from the too small sample size in planned domains.

3.1.1. Special cases of balanced samples

Except multistage sampling, almost all the other sampling designs are particular cases of balanced sampling:

1. Example. Sampling with a fixed sample size. It is a balance sample if the only balancing variable is $\pi_k(k)$. The balancing equations given in (3.2) become

$$\sum_{k\in s(t)} \frac{\pi_k(t)}{\pi_k(t)} = \sum_{k\in s(t)} 1 = \sum_{k\in U} \pi_k(t) = n(t),$$

   which means that the sample size must be fixed.

2. Example. Stratified simple random sampling without replacement. It is a balance sample if balancing variables are the indicator variables of the strata:

$$\delta_{h,k} = \begin{cases} 1, & \text{if } k \in U_h, \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

   Here $h = 1, \ldots, H$. Under a stratified design, the Horvitz-Thompson estimators of the sizes of the strata exactly equal the sizes of the strata, which is a property of balancing on the indicator variables of the strata. Indeed, since the inclusion probabilities in stratum h are $\pi_k(t) = n_h(t)/N_h$, $k \in U_h$, the balancing equations become

$$\sum_{k\in s(t)} \frac{N_h \delta_{h,k}}{n_h(t)} = \sum_{k\in U} \delta_{h,k} = N_h, \quad h = 1, \ldots, H,$$

and are exactly satisfied.

3. Example. Another interesting special case of balanced sampling occurs when a constant is used as a balancing variable. If $z_k(t) = 1$ for all $k \in U$, the balancing equations become

$$\sum_{k \in s(t)} \frac{1}{\pi_k(t)} = \sum_{k \in U} 1 = N.$$

   Actually, the left part of this equation is the Horvitz-Thompson estimator of N. This means that, if a constant is used as a balancing variable, the estimated population size matches the known size N, which is far from being a given when the statistical units are selected with unequal inclusion probabilities.

In fact, balanced sampling is a more general method of sampling that includes almost all the other methods.

3.1.2. The choice of balanced variables

There are several recommendations how to choose balancing variables (Tillé (2011)):

1. The main recommendation is to choose balancing variables that are closely correlated to the study variable or variables.
2. Not choose too many balancing variables because, accuracy no longer improves with a large number of variables and the instability of the variance estimator increases with each additional variable.
3. The auxiliary variables should not be too correlated amongst themselves.

In many cases, the balancing variables contain measurement errors. For example, in most registers, missing values can obviously occur and auxiliary variables are often corrected by a method of imputation. Indeed, the gain in efficiency only depends on the correlation between the balancing variables and the study variable. This correlation is rarely affected by errors in the balancing variables.

Several auxiliary variables can be used to improve small domain estimates. To ensure that a domain D is not empty, it is possible to add such auxiliary variable:

$$z_k(t) = \begin{cases} \pi_k(t), & \text{if } k \in U^{(D)}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

This implies that the number of sampled units that belong to D domain is equal to

$$n^{(D)}(t) = \sum_{k \in U} z_k(t) = \sum_{k \in U^{(D)}} \pi_k(t).$$

3.1.3. Balanced sample and the use of auxiliary information in estimation stage

A balanced sample increases the accuracy of the Horvitz-Thompson estimator, thus it is possible to think, that there is no need to use auxiliary information in estimation stage. Indeed, in many cases, at the estimation stage, more auxiliary variables are often available, thus the use of GREG-type estimator might improve estimators even more.

Generally, it is recommended to incorporated and the same auxiliary variables, us it were used as balanced variables, because if only new variables will be used in estimation stage, the effect of balancing can be lost.

There is, however, one case where only new variables can be used at estimation stage: when the balancing variables are no longer correlated to the study variable. This can occur when the balancing and the new variables are the same variables measured at different moments, and the new variables are more recent.

3.1.4. Cube method

One of the methods how to select a balanced sample is to use Cube method. The algorithm of the cube method was proposed by Deville and Tillé (1998) and the method was published by Deville and 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.

The name of the method comes from the geometric representation of a sampling design. Indeed, a sample may be represented by a vector of sample indicators $\mathbf{S}(t) = (S_1(t), S_2(t), \ldots, S_N(t))$ where $S_k(t)$ takes value 1 if $k \in \mathbf{S}(t)$ and 0 if not. A sample may thus be viewed as a vertex of an N-cube.

The algorithm of the cube method enables us to run a function with two arguments: the vector of inclusion probabilities and the matrix of balancing variables. It is based on a random transformation of the vector of inclusion probabilities until a sample is obtained such that:

1. The inclusion probabilities are exactly satisfied.

2. The balancing equations are satisfied to the furthest extent possible.

The Cube method is divided into two phases: the flight phase and the landing phase. The flight phase is a random walk that begins at the vector of inclusion probabilities and remains in the intersection of the cube and the constraint subspace. This random walk stops at a vertex of the intersection of the cube and the constraint subspace. At the end of the flight phase, if a sample is not obtained, the landing phase entails in selecting a sample that is as close as possible to the constraint subspace.

In some cases, it is interesting to balance on auxiliary variables in subgroups, domains or strata. An interesting procedure described in Chauvet (2009) consists of separately running the flight phase in each stratum. A rounding problem will then occur in each stratum. These rounding problems can then be merged and a flight phase can be run again on the whole population. Finally, the landing phase is applied only to the whole population. This procedure enables us to roughly satisfy the balancing equations in each strata without cumulating the rounding problems.

Let us look at this algorithm more precisely. Thus, at each step in the flight phase, it is randomly chosen to either select or permanently discard one of the population unit. At the end of the flight phase, in each stratum U_h there is a vector $\pi^*_h(t) = (\pi^*_k(t))_{k\in U_h} \in [0,1]^N$, that satisfies the following conditions:

$$E(\pi^*_h(t)) = \pi_h(t), \tag{3.5}$$

$$\sum_{k\in U_h} \frac{\mathbf{z}_k(t)}{\pi_k(t)}\pi^*_k(t) = \sum_{k\in U_h} \mathbf{z}_k(t), \tag{3.6}$$

$$Card\{k \in U_h;\, 0 < \pi^*_k(t) < 1\} \leq P, \tag{3.7}$$

where E denotes the expectation for the sampling method used in the flight phase. The vector $\pi^*_h(t)$ gives the outcome of the flight phase: $\pi^*_k(t)$ is 1 if unit k is selected in time t, 0 if it is rejected, and between 0 and 1 only if the decision has not been made for unit k after the flight phase.

Equations (3.5) and (3.6) ensure that the inclusion probabilities and balancing constraints are maintained perfectly at the end of the flight phase. Equation (3.7) ensures that a decision remains to be made for no more than P individuals in each stratum, U_h, where P is the number of balancing variables.

The flight phase ends when the balancing constraints can no longer be exactly satisfied. The landing phase consists in defining, conditionally on the outcome of the flight phase, an optimal sampling design defined on the remaining population $V(t)$. This design is optimal in that it makes it possible to complete the sampling while minimizing the variance, conditionally on the outcome of the flight phase, of the Horvitz-Thompson estimator of the balancing variables. The remaining units are sampled, conditionally on the outcome of the flight phase, with inclusion probabilities $\pi^*_k(t)_{k\in V(t)}$ so that the units' unconditional inclusion probabilities $\pi_k(t)_{k\in V(t)}$ are maintained exactly.

In the case of stratified balanced sampling, the variance of Horvitz-Thompson estimator is

$$var_p(\widehat{\underline{TOT}}_{HT,bal}(t)) \simeq \sum_{h=1}^{H} \sum_{k\in U_h} \frac{b_k(t)}{\pi_k^2(t)}(y_k(t) - \beta_h(t)\mathbf{z}_k(t))^2, \tag{3.8}$$

where

$$\beta_h(t) = \Big(\sum_{k\in U_h} b_k(t)\frac{\mathbf{z}_l(t)}{\pi_k(t)}\frac{\mathbf{z}'_l(t)}{\pi_k(t)}\Big)^{-1} \sum_{k\in U_h} b_k(t)\frac{\mathbf{z}_l(t)}{\pi_k(t)}\frac{y_l(t)}{\pi_k(t)}. \tag{3.9}$$

Deville and Tillé (2005) offer several approximations for the $b_k(t)$. The simplest is $b_k(t) = \pi_k(t)(1 - \pi_k(t))$. The variance of the Horvitz-Thompson estimator will be small if, in each stratum, study variable y is well explained by balancing variables $\mathbf{z}(t)$.

Using equation (3.8) it is possible to write the variance of Horvitz-Thompson estimator for the domain total:

$$var_p(\widehat{\underline{TOT}}^{(d)}_{HT,bal}(t)) \simeq \sum_{h=1}^{H} \sum_{k\in U_h\setminus U^{(d)}} \frac{b_k(t)}{\pi_k^2(t)}(y_k(t) - \beta_h(t)\mathbf{z}_k(t))^2, \tag{3.10}$$

where $d = 1, \ldots D$.

If sample set s is selected from U in accordance with the stratified balanced sampling procedure described above, sampling will be balanced in each stratum as long as the landing phase affects a small number of individuals relative to the sample size. Specifically, equation (3.7) shows that the number of balancing variables must be small relative to the sample allocation in each stratum. In some cases, that constraint cannot be satisfied. The population is often partitioned into very small groups to make the results more relevant, which means decreasing the number of units selected in each stratum. In this case balanced sampling will be only very approximate.

There is a SAS/IML version done by Chauvet and Tillé (2005) which is available on the University of Neuchatel Web site. Also there is possibility to use cube method in R language (Tillé and Matei (2007)). These software programs are free, available over the internet and are easy to use.

3.2. Model-based sample design

There are two cases, when the balanced sample might not work as good at it is expected:

1. When the balancing variables are no longer correlated to the study variable.
2. There are a lot of small domains.

Thus let us discussed and some other sample methods. Let $y_k(t)$ and $\mathbf{z}_k(t)$, $k = 1, ..., N$, be the realizations of random variables $\underline{y}_k(t)$ and $\underline{\mathbf{z}}_k(t)$ of the superpopulation model $\mathcal{M}$:

$$\underline{y}_k(t) = \underline{\mathbf{z}}_k(t)'\beta(t) + \varepsilon_k(t), \quad k \in U. \tag{3.11}$$

Here $\underline{\mathbf{z}}_k(t) = (1, \underline{z}_{1,k}(t), ..., \underline{z}_{P,k}(t))'$ are not random, $\beta(t) = (\beta_0, \beta_1, ..., \beta_P)'$ are the model coe cients, $E_{\mathcal{M}}(\varepsilon_k(t)) = 0$, $var_{\mathcal{M}}(\varepsilon_k(t)) = \nu_k^2\sigma^2$, for all $k \in U$, and $cov_{\mathcal{M}}(\varepsilon_k(t), \varepsilon_l(t)) = 0$, when $k \neq l$. Also there is made an assumption that ν_k are known and $\sum_{k \in U} \nu_k = N$. Then Nedyalkova, Tille (2008) showed, that under superpopulation model (3.11), an optimal model-assisted strategy consists of using inclusion probabilities that are proportional to ν_k, selecting the sample by means of a balanced sampling design on $\mathbf{z}_k$, and using Horvitz-Thompson estimator (see 2.2.1).

In this research more general superpopulation model than (3.11) is considered:

$$\begin{aligned}\underline{y}_k(t) =& \beta_{0,g(k)}(t) + v_{0,k}(t) + \sum_{p=1}^{P} [\beta_{p,g(k)}(t) + v_{p,k}(t)]\underline{z}_{p,k}(t) + \\ &+ \sum_{i=1}^{m} \alpha_{i,g(k)}\mu_i(t) + \varepsilon_k(t), \quad k \in U.\end{aligned} \tag{3.12}$$

This model has the same assumptions as model (2.57), just instead of auxiliary variable $\underline{\mathbf{x}}_k(t)$, the variable $\underline{\mathbf{z}}_k(t)$ is used. That is done, because the set of covariates 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.

In this case the construction of balanced sample for all auxiliary variables

$\underline{\mathbf{z}}_k(t)$ may not be reasonable from the practical point of view since it requires to select in advance informative auxiliary variables. It might be enough to use a model-based design. Thus the improvement of the estimation might be done using GREG-type estimator (see section 2.2.2) instead of Horvitz-Thompson estimator.

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 best superpopulation model is fitted to the available auxiliary information AI (2.56 equation).

In the second step the prediction errors (residuals)

$$\hat{\varepsilon}_k(t) = \hat{y}_k(t) - y_k(t), \quad t \in \mathcal{T}_k \subset \{1, 2, \ldots, T\}, \tag{3.13}$$

are calculated and the variance of prediction error $\nu_k^2\sigma^2$ is estimated.

Finally, in the third step the (approximately) optimal sample design $p(.)$ based on the variance of prediction error is chosen. Here really it possible to use stratified simple random sample (where strata are constructed using the variance of prediction error) or even probability proportional to size sampling (where size variable is approximately equal to the variance of prediction error).

3.3. The summary of the third chapter

1. To improve estimation strategy, a balance sample or model-based sample design can be used.

2. A balance sample is such sample where the total's of balanced variables are estimated without error. In the section 3.1.1. it is showed, that some common used sample designs are balance.

3. To get a good balance sample it is recommended that balancing variables should be closely correlated with the study variable, but not amongst themselves. Also the number of balance variables must not be large, because, accuracy no longer improves with a large number of variables and the instability of the variance estimator increases with each additional variable.

4. One of the way to get a balance sample is to use cube method.

5. A use of balance sample might not give desirable improvement of accuracies if here are a lot of small domains or the balancing variables are no longer correlated to the study variable.

6. The proposed model-based sample consists of three steps: model construction and estimation of it's coefficients, estimation of the variance of the prediction error and Construction of the sample design.

4

Monte Carlo studies using real data

4.1. Study population

For the simulation experiment, a real population from Statistics Lithuania is used. Enterprises which are responsible for education are taken as the finite population. Information about these enterprises 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). Thus, 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.

Such population is chosen, because there are a lot of surveys in Lithuania and other countries, that the relation between study variable and auxiliary variables are similar. Furthermore the data are not homogeneous, thus the separate cases are investigate in separate domains.

From the real population (indicated by the number $R = 0$), two more population where constructed by generating different response rate. The response rates of 80% and 70% were generated for the first ($R = 1$) and the second ($R = 2$) populations, respectively. These rates represents the response rate in the survey (actually the response rate depends on county, number of employees and specification).

4.2. Accuracy measures

Two accuracy measures were applied to compare the performance of the different estimators for $M = 1000$ simulation. They are the absolute relative bias

$$ARB(\widehat{TOT}_{*}^{(d)}(t)) = \frac{\left|\frac{1}{M}\sum_{m=1}^{M}\widehat{TOT}_{*,m}^{(d)}(t) - TOT^{(d)}(t)\right|}{TOT^{(d)}(t)} \tag{4.1}$$

and the relative root means square error

$$RRMSE(\widehat{TOT}_{*}^{(d)}(t)) = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{TOT}_{*,m}^{(d)}(t) - TOT^{(d)}(t)\right)^2}}{TOT^{(d)}(t)}. \tag{4.2}$$

Here $\widehat{TOT}_{*,m}^{(d)}(t)$ is the predicted value of the total from m-th simulation in domain d using $*$ estimator (what estimators are used in this research it is written in section 4.3.1) and the $TOT^{(d)}(t)$ refers to the true population in the same domain at the same time t.

There are 120 domains of interest, so for the better comparison these domains where grouped in three groups g_i, $i = 1, \ldots, 3$ (see table 4.1).

Table 4.1. Groups of the domains

Group	Number of domain in group	Number of sampled enterprises in the domain, (min - max)
Small	48	$0 - 9$
Medium	40	$10 - 29$
Large	32	> 30
Total	120	

A mean of absolute relative bias and mean of relative root means square error in each group were calculated:

$$MARB = \frac{\sum_{d \in g_i} ARB(\widehat{TOT}_*^{(d)}(t))}{\sum_{d \in g_i} 1}; \tag{4.3}$$

$$MRRMSE = \frac{\sum_{d \in g_i} RRMSE(\widehat{TOT}_*^{(d)}(t))}{\sum_{d \in g_i} 1}. \tag{4.4}$$

4.3. Monte Carlo study I: Comparison of small area estimators in the case of nonresponse

Using this simulation several things were compared:

1. The performance of different types of estimators using different sample designs (see section 4.3.2).
2. The performance of model assisted and model-based estimators under different superpopulation models(see sections 4.3.3 and 4.3.4).
3. The performance of different nonresponse methods using different estimators and response rates (see section 4.3.5).

4.3.1. Sample designs, nonresponse adjustment and estimators

Different types of estimators and nonresponse adjustment methods are applied to 1000 samples generated by each of the following three sample designs for all three populations:

1. SRS – simple random sample of 300 elements. Here the domain size is unplanned.
2. SSRS – simple random stratified sample. Here population is divided into three strata by the size of enterpriser. The number of selected enterprises from each strata are presented in table 4.2.
3. SπPS – stratified sampling proportional to size. Here the strata and the number of selected enterprises in each strata are the same as for SSRS (see table 4.2), just the probability to select enterpriser in to the sample is proportional to the number of employees in the current enterprise.

Table 4.2. Number of enterprises in the strata

Strata	Number of enterprises in population	Number of enterprises in the sample
1	561	160
2	114	70
3	75	70
Total	750	300

Three types of estimators were chosen for estimation: Horvitz-Thompson, GREG-type and Model-based. The notation of the estimator is constructed in this way: $E_{\mathcal{M}_c} - LLR$. The $E \in \{HT, GREG, MB\}$ notate which estimator is used, $\mathcal{M}_c$, $c \in \{1, 2, \ldots, 10\}$ notate what model is used as superpopulation model, two letters, $LL \in \{WC, LR, RD, NN, CR, DR\}$, notate the nonresponse adjustment method and number $R \in \{0, 1, 2\}$, notate which population is used.

Thus 10 different superpopulation models were used as a assisted tool for GREG-type estimator, or as the model for model-based estimator (see table 4.3).

The first model (see table 4.3) is based on the model with two auxiliary variables. The model's coefficients are the same for all domains but differs from time to time. Thus the model's coefficients in time t are estimated using just response sample data of the same time moment. The estimators, which use this model us superpopulation model are indirect, because the model's coefficients for domain are estimated using data not only from the same domain, but and from the others.

The difference between first and second model is just in the intercept. For the first model it is the same for whole population, thus for the second model it is separate for the groups. Here the groups divide enterprises by the size of enterpriser and region.

In the model $\mathcal{M}_3$ the difference between groups is made by adding random intercept $v_{0,g(k)}(t)$. This model, as and the previous two, has different model's coefficients from time to time, thus they can be calculated using just the single sample data from the same time t as the parameter of interest.

The same property has and $\mathcal{M}_4$ model, but the coefficients of this model differs and between groups. It means that for some small groups model's coefficients are estimated using small number of enterprises.

The next 4 models $\mathcal{M}_5$, $\mathcal{M}_6$, $\mathcal{M}_7$ and $\mathcal{M}_8$ are panel type models, because the model's coefficients do not depend on time, thus they might be estimated using huge amount of auxiliary information which is possible to get even be-

Table 4.3. Models used in estimation stage

Notation	Model
$\mathcal{M}_1$	$\underline{y}_k(t) \;=\; \beta_0(t)+\beta_1(t)\underline{x}_{1,k}(t)+\beta_2(t)\underline{x}_{2,k}(t)+$ $+\varepsilon_k(t),\ k\in U$
$\mathcal{M}_2$	$\underline{y}_k(t) \;=\; \beta_{0,g(k)}(t)+\beta_1(t)\underline{x}_{1,k}(t)+\beta_2(t)\underline{x}_{2,k}(t)+$ $+\varepsilon_k(t),\ k\in U$
$\mathcal{M}_3$	$\underline{y}_k(t) \;=\; \beta_0(t)+v_{0,g(k)}(t)+\beta_1(t)\underline{x}_{1,k}(t)+\beta_2(t)\underline{x}_{2,k}(t)+$ $+\varepsilon_k(t),\ k\in U$
$\mathcal{M}_4$	$\underline{y}_k(t) \;=\; \beta_{0,g(k)}(t)+\beta_{1,g(k)}(t)\underline{x}_{1,k}(t)+\beta_{2,g(k)}(t)\underline{x}_{2,k}(t)+$ $+\varepsilon_k(t),\ k\in U$
$\mathcal{M}_5$	$\underline{y}_k(t) \;=\; \beta_0+\beta_1\underline{x}_{1,k}(t)+\beta_2\underline{x}_{2,k}(t)+$ $+\varepsilon_k(t),\ k\in U$
$\mathcal{M}_6$	$\underline{y}_k(t) \;=\; \beta_{0,g(k)}+\beta_1\underline{x}_{1,k}(t)+\beta_2\underline{x}_{2,k}(t)+$ $+\varepsilon_k(t),\ k\in U$
$\mathcal{M}_7$	$\underline{y}_k(t) \;=\; \beta_0+v_{0,g(k)}+\beta_1\underline{x}_{1,k}(t)+\beta_2\underline{x}_{2,k}(t)+$ $+\varepsilon_k(t),\ k\in U$
$\mathcal{M}_8$	$\underline{y}_k(t) \;=\; \beta_{0,g(k)}+\beta_{1,g(k)}\underline{x}_{1,k}(t)+\beta_{2,g(k)}\underline{x}_{2,k}(t)+$ $+\varepsilon_k(t),\ k\in U$
$\mathcal{M}_9$	$\underline{y}_k(t) \;=\; \beta_{0,g(k)}+\beta_1\underline{x}_{1,k}(t)+\beta_{2,g_1(k)}\underline{x}_{2,k}(t)+$ $+\sum_{j=1}^{3}(a_j+b_j\underline{x}_{2,k}(t))s_j+\varepsilon_k(t),\ k\in U$
$\mathcal{M}_{10}$	$\underline{y}_k(t) \;=\; \beta_{0,g(k)}+v_{0,g_1(k)}+\beta_1\underline{x}_{1,k}(t)+\beta_{2,g(k)}\underline{x}_{2,k}(t)+$ $+\sum_{j=1}^{3}(a_j+b_j\underline{x}_{2,k}(t))s_j+\varepsilon_k(t),\ k\in U$

fore current sample is selected.

The last two models $\mathcal{M}_9$ and $\mathcal{M}_{10}$ also incorporates the linear trend and seasonal components and for model $\mathcal{M}_{10}$ random intercept for the groups also is included. In this case the groups are divided by the regions.

Six different nonresponse adjustment method are used to estimate parameters of interest in two populations with different response rate (80% for population $R = 1$ and 70% for population $R = 2$). The differences between nonresponse adjustment methods are denoted by adding two letters, $LL \in \{WC, LR, RD, NN, CR, DR\}$, at the end of the estimate's name. The meaning of these abbreviations is described below. The weighting-class method (WC) and the logistic regression model (LR) are applied to estimate the response probability. Also, the performance of different imputation methods (random donor (RD), nearest neighbors (NN), regression imputation using the common model (CR)(Lehtonen, Särndal and Veijanen (2003)) and regression imputation using the model with domain-intercepts (DR) (Lehtonen, Särndal and Veijanen (2003))) is investigated. For weighting-class, random donor and nearest neighbors methods units are grouped by the number of employees and specification.

4.3.2. Design-based estimators vs model-based estimators

From the survey sampling theory it is known, that design-based estimators are unbiased (HT case) or approximately unbiased (GREG case). As for the model-based estimators – the variance in usually smaller than for design-based estimators, but they are design-bias.

Thus in this research these properties of the design-based and model-based estimators are verified in the case of small area estimation (see table 4.4).

The results showed what was expected: design-based estimators are approximately unbiased even when the sample size is small. The model-based estimator is bias and in some cases the bias is so large, that RRMSE is bigger then for design-based estimators even if it is clear that the variance of model-based estimator is significantly smaller than for design based estimator.

Here for the superpopulation model, a common model is chosen. It seems, that it is not the best model for this population, thus the different models are compared in the next two sections.

Table 4.4. Performance of different types of estimators using different sample designs

	Domain sample size classes					
Estimator	Small $0-9$		Medium $10-29$		Large $30-\ldots$	
	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$
	Simple random sampling (SRS)					
HT	2.0	56.4	3.2	38.6	2.1	23.6
$GREG_{\mathcal{M}_1}$	1.4	50.9	2.2	17.9	0.9	10.3
$MB_{\mathcal{M}_1}$	48.5	55.6	18.4	24.1	5.5	11.2
	Stratified simple random sampling (SSRS)					
HT	2.4	31.6	0.7	15.2	0.4	8.8
$GREG_{\mathcal{M}_1}$	1.8	24.3	0.7	14.3	0.2	5.7
$MB_{\mathcal{M}_1}$	23.5	26.7	12.4	17.7	2.6	5.7
	Stratified sampling proportional to size (SπPS)					
HT	2.3	52.4	0.7	18.8	0.4	8.4
$GREG_{\mathcal{M}_1}$	2.1	36.4	1.1	17.0	0.4	7.4
$MB_{\mathcal{M}_1}$	22.5	26.9	9.6	13.7	2.8	5.5

4.3.3. Models from the current sample vs panel type models

The idea of comparison of the models from the current sample and panel-type model is based on these assumptions:

1. To estimate reliable model's coefficient from the current sample might be impossible especially when the sample size in domain is small.
2. The use of panel-type models where model's coefficients do not depend on time increase the amount of data from which the model's coefficients are estimated, thus the estimates of model's coefficients might be more reliable.

The results are showed in table 4.5 and table 4.6. The same sample design and superpopulation models are used in both tables, just in table 4.5 the performance of Greg-type estimator is investigated while in table 4.6 – Model-based estimator.

The results in these tables are group by the model's coefficients specification.

In the first group model's with the same coefficients for the whole population are compared. Here for the model $\mathcal{M}_1$ coefficients depends on time, thus the estimation of them can be done just using current sample data. As for the model $\mathcal{M}_5$ coefficient are the same and for the whole population and from time

Table 4.5. Performance of different models when GREG-type estimator is used. A stratified simple random sample case

	Domain sample size classes					
Estimator	Small $0-9$		Medium $10-29$		Large $30-\dots$	
	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$
	Models where model's coefficients are the same for whole population					
$\mathcal{M}_1$	1.8	24.3	0.7	14.3	0.2	5.7
$\mathcal{M}_5$	1.7	25.0	0.7	14.9	0.3	5.9
	Models where fixed intercepts are different between groups					
$\mathcal{M}_2$	1.7	26.2	0.6	14.2	0.3	5.8
$\mathcal{M}_6$	1.6	24.1	0.6	14.0	0.2	5.7
	Models where random intercepts are different between groups					
$\mathcal{M}_3$	1.7	26.3	0.6	14.2	0.3	5.9
$\mathcal{M}_7$	1.6	24.1	0.6	14.0	0.2	5.7
	Models where model's coefficients are different between groups					
$\mathcal{M}_4$	1.1	23.0	0.4	14.4	0.3	5.7
$\mathcal{M}_8$	1.1	22.9	0.7	13.3	0.2	5.7
$\mathcal{M}_9$	0.8	20.0	0.7	12.7	0.2	5.5
$\mathcal{M}_{10}$	0.8	20.0	0.7	12.7	0.2	5.5

to time. This model is perfect if the differences between enterprises can be expressed just through the auxiliary information. In this case the results showed, that the income of enterpriser depends not only on the number of employees and VAT, but and on some other information (the results for the model $\mathcal{M}_5$ is worse than for the model $\mathcal{M}_1$ which also depends on time).

In the second group and third groups models with different intercepts for the groups are investigates. The differences between groups are made using different fixed-effect intercepts (second group of results) or using random-effect intercepts (third group of results). The results of these groups already showed the the panel type models give better results, thus the use of different intercept in the panel type model gives larger improvement than the use of the model where coefficients depend on time (see model $\mathcal{M}_6$ and $\mathcal{M}_1$ in table 4.5 or model $\mathcal{M}_7$ and $\mathcal{M}_1$ in table 4.6).

In the last group of results models with different coefficient in the groups are compared. For the models $\mathcal{M}_9$ and $\mathcal{M}_{10}$ the linear trend and seasonal components are incorporated and for the model $\mathcal{M}_{10}$ random-effect for intercept is also included. The results showed, that panel type model with time components (trend and seasonal components) is the best choice for both – design-based and

Table 4.6. Performance of different models when Model-based estimator is used. A stratified simple random sample case

	Domain sample size classes					
Estimator	Small $0-9$		Medium $10-29$		Large $30-\ldots$	
	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$
	Models where model's coefficients are the same for whole population					
$\mathcal{M}_1$	23.5	26.7	12.4	17.7	2.6	5.7
$\mathcal{M}_5$	25.5	30.1	13.1	18.5	2.9	5.9
	Models where fixed intercepts are different between groups					
$\mathcal{M}_2$	22.9	27.4	12.9	18.6	2.7	5.4
$\mathcal{M}_6$	21.2	25.3	12.5	18.0	2.6	5.3
	Models where random intercepts are different between groups					
$\mathcal{M}_3$	22.5	26.4	11.9	17.2	2.8	5.5
$\mathcal{M}_7$	21.0	25.0	11.7	16.9	2.7	5.3
	Models where model's coefficients are different between groups					
$\mathcal{M}_4$	25.1	30.4	8.6	15.5	2.3	5.7
$\mathcal{M}_8$	20.2	24.3	7.7	14.3	2.2	5.4
$\mathcal{M}_9$	19.4	23.6	7.6	14.2	2.2	5.2
$\mathcal{M}_{10}$	19.2	23.3	7.5	14.0	2.2	5.1

model-based estimators.

4.3.4. Models with random effect vs models without random effect

The choice between models with random-effect and without random-effect is investigated in this section. Such kind of investigation is made because in Lehtonen, Särndal and Veijanen (2003) paper it was made an conclusion, that for design-based model assisted estimators better to use models without random effects and for model-based estimators – models with random-effects. Still in this paper models of the current sample are investigated and in this research it was already showed, that it is better to use the panel type model instead of the model based on the data from the current sample.

For the comparison random-effect model and model without random effect a simple random sample (SRS) design is chosen. The results for Greg type estimator are presented in table 4.7 while for the Model-based estimator – in table 4.8. All these results in both tables are grouped in to three groups.

In the first group models where model's coefficients are estimated us-

Table 4.7. Performance of different models when GREG-type estimator is used. A simple random sample case

Estimator	Domain sample size classes					
	Small $0-9$		Medium $10-29$		Large $30-\ldots$	
	$MABR, \%$	$MRRMSE, \%$	$MABR, \%$	$MRRMSE, \%$	$MABR, \%$	$MRRMSE, \%$
	Models where model's coefficients are estimated using current sample data					
$\mathcal{M}_2$	2.0	53.7	2.1	18.5	0.7	10.6
$\mathcal{M}_3$	2.0	53.8	2.2	18.7	0.7	10.6
	Models where model's coefficients are estimated using data from the past					
$\mathcal{M}_6$	1.8	48.7	1.7	17.5	0.7	10.4
$\mathcal{M}_7$	1.8	48.7	1.7	17.5	0.7	10.4
	Panel type models with linear trend and seasonal components					
$\mathcal{M}_9$	1.4	46.4	1.6	16.6	0.6	10.3
$\mathcal{M}_{10}$	1.5	46.3	1.6	16.6	0.6	10.3

ing current sample design are compared. Here the model's coefficients are the same for the whole population except the intercept. For the model $\mathcal{M}_2$ the intercept is fixed and differs between groups and for the model $\mathcal{M}_3$ the intercept is random for the same groups as in model $\mathcal{M}_2$.

Table 4.8. Performance of different models when model-based estimator is used. A simple random sample case

Estimator	Domain sample size classes					
	Small $0-9$		Medium $10-29$		Large $30-\ldots$	
	$MABR, \%$	$MRRMSE, \%$	$MABR, \%$	$MRRMSE, \%$	$MABR, \%$	$MRRMSE, \%$
	Models where model's coefficients are estimated using current sample data					
$\mathcal{M}_2$	47.0	56.3	15.1	20.5	5.4	10.3
$\mathcal{M}_3$	48.5	55.6	14.9	19.9	5.3	10.2
	Models where model's coefficients are estimated using data from the past					
$\mathcal{M}_6$	42.9	50.6	14.0	19.3	5.0	9.6
$\mathcal{M}_7$	42.7	50.1	13.8	19.0	5.0	9.5
	Panel type models with linear trend and seasonal components					
$\mathcal{M}_9$	41.4	49.6	13.9	19.0	4.9	9.2
$\mathcal{M}_{10}$	41.3	49.4	13.7	18.9	4.9	9.1

In the second group the panel type models are compared. For these models coefficients do not depend on time, thus they are estimated using data from the past. For both model coefficients are the same for the whole population except intercept (for the model $\mathcal{M}_6$ the intercept differs between groups but it is fixed and for the model $\mathcal{M}_7$ the intercept is random for the same groups as in model $\mathcal{M}_6$).

In the third group the panel type model with time components are compared. Here the model's coefficients are different between groups (size of enterpriser and region) and in model $\mathcal{M}_{10}$ the random intercept is incorporated for larger groups (region).

The results for the GREG-type estimator are so similar that it is no possible to say which model (with random effect or without it) is better. Therefore the performance of the hypothesis testing of equality of two variances is taken as an additional criterion for the comparison. Here the results showed that in most of the cases the difference between variances if not significant at 10% level. Just for the 20% of domains (most of them are small ones) a model without random effect give better results then the use of the model with random effect.

The results for the model-based estimator is different. Here models with random-effect (compared with models without random-effect) reduce the $MARB$ and $MRRMSE$.

4.3.5. Weighting methods vs imputation methods

A comparison between different nonresponse adjustment methods is made in this section.

Weighting methods, imputation using donors and imputation using models are compared at first (see table 4.9). Here the population with 80% response rate is used. For the GREG-type estimator model $\mathcal{M}_9$ is used as the assisted tool and for the model-based estimator – $\mathcal{M}_{10}$. They are chosen as the best models for these estimators (see sections 4.3.3 and 4.3.4).

The results from the table 4.9 showed, that for the both estimators the imputation using donor increase the $MARB$ and $MRRMSE$ more than using other methods. The imputation using models increase $RRMSE$ more than weighting methods, thus in this research the best way to adjust nonresponse is to use re-weighting, when inclusion probabilities are estimated using logistic regression model.

The comparison between estimators when different response rate occurs is the second task. Here the results are presented in tables 4.9 and 4.10. It shows, that when response rate decreases the bias of the GREG-type estimator increases quicker than the bias of the model-based estimator. The results in table

Table 4.9. Comparison of different nonresponse adjustments methods when response rate is 80%. A stratified simple random sample case

	Domain sample size classes					
Estimator	Small $0-9$		Medium $10-29$		Large $30-...$	
	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$
	GREG-type estimator					
$GREG_{\mathcal{M}_9}-WC1$	5.1	22.7	3.3	14.0	1.9	6.4
$GREG_{\mathcal{M}_9}-LR1$	4.8	21.5	3.1	13.6	1.8	6.2
$GREG_{\mathcal{M}_9}-RD1$	9.5	44.3	6.5	25.9	3.7	10.2
$GREG_{\mathcal{M}_9}-NN1$	8.3	37.6	4.9	21.2	3.1	9.5
$GREG_{\mathcal{M}_9}-CR1$	4.9	22.3	3.2	13.9	1.8	6.3
$GREG_{\mathcal{M}_9}-DR1$	5.0	22.6	3.3	13.9	1.8	6.4
	Model-based estimator					
$MB_{\mathcal{M}_{10}}-WC1$	19.3	23.5	7.6	14.1	2.2	5.2
$MB_{\mathcal{M}_{10}}-LR1$	19.2	23.5	7.5	14.0	2.2	5.1
$MB_{\mathcal{M}_{10}}-RD1$	25.3	29.8	9.9	18.2	4.3	7.8
$MB_{\mathcal{M}_{10}}-NN1$	24.7	28.4	9.6	17.9	4.1	7.5
$MB_{\mathcal{M}_{10}}-CR1$	19.3	24.8	7.6	15.8	2.3	6.0
$MB_{\mathcal{M}_{10}}-DR1$	19.2	24.7	7.7	15.6	2.3	5.8

Table 4.10. Comparison of different nonresponse adjustments methods when response rate is 70%. A stratified simple random sample case

	Domain sample size classes					
Estimator	Small $0-9$		Medium $10-29$		Large $30-...$	
	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$
	GREG-type estimator					
$GREG_{\mathcal{M}_9}-WC1$	7.3	25.6	4.8	17.2	2.3	8.7
$GREG_{\mathcal{M}_9}-LR1$	6.9	25.4	4.7	16.8	2.2	8.6
$GREG_{\mathcal{M}_9}-RD1$	15.7	54.5	8.6	32.5	4.4	13.2
$GREG_{\mathcal{M}_9}-NN1$	12.1	45.6	7.7	26.3	3.6	11.5
$GREG_{\mathcal{M}_9}-CR1$	6.9	28.2	5.1	16.9	2.3	9.2
$GREG_{\mathcal{M}_9}-DR1$	7.1	27.8	5.2	16.8	2.2	9.0
	Model-based estimator					
$MB_{\mathcal{M}_{10}}-WC1$	19.7	24.6	7.9	14.8	2.2	5.6
$MB_{\mathcal{M}_{10}}-LR1$	19.7	24.4	7.5	14.6	2.3	5.5
$MB_{\mathcal{M}_{10}}-RD1$	27.0	32.2	10.9	20.1	4.6	8.8
$MB_{\mathcal{M}_{10}}-NN1$	26.4	30.5	10.4	19.4	4.4	8.5
$MB_{\mathcal{M}_{10}}-CR1$	19.8	28.0	7.8	17.8	2.4	6.4
$MB_{\mathcal{M}_{10}}-DR1$	19.9	27.6	7.8	17.5	2.3	6.2

4.9 show that event in the GREG-type estimator becomes bias when nonresponse occurs, it still has smaller $RRMSE$ than the model-based estimator. As for the table 4.10 the results are opposite: model-based estimator perform better results than GREG-type estimator (except for the bias of small domains). Thus when the response rate decreases more than some point (in this research more than 70%) it is better to use model-based estimator, while till that point GREG-type estimator gives better results than model-based estimator.

4.4. Monte Carlo study II: Searching for the optimal strategy when panel type data are used

In this section several estimation strategies are used to answer the following problems: what type of model, sample design and estimator should be used in small area estimation.

4.4.1. Estimation strategies

For the simulation experiment, the same population as in section 4.1 is considered. The study variable $y_k(t)$, the auxiliary variables $x_{j,k}$, $j = 1, ..., 15$ and parameter of interest are also the same.

The purpose of this research is to find an optimal strategy (pair of sample design and estimator) for small are estimation when the panel type data are used. Thus this purpose can be divided into these steps:

1. To improve strategies what are showed in chapter 4.3 a balance sample is selected, where first order inclusion probabilities are defined in the same way as for SRS, SSRS and SπPS designs. For these sample designs different number of the balanced variables are used.
2. Compare the results of best strategies from the previous step when nonresponse occurs.
3. To improve strategies by using model-based sample design.

The results of the each step are presented in sections 4.4.2, 4.4.3 and 4.5 respectively.

4.4.2. Comparison of strategies when balanced sample is used

A balanced samples are selected using different inclusion probabilities and different number of balanced variables.

Three different methods are used to estimate inclusion probabilities:

1. The inclusion probabilities are the same for all units. In this case inclusion probabilities are the same as using SRS design, thus the notation of a balanced sample when inclusion probabilities are the same is SRS.
2. The inclusion probabilities are the same for the units from the same strata. In this case inclusion probabilities are the same as using SSRS design, thus the notation of a such balanced sample is SSRS.
3. The inclusion probabilities are proportional to the size variable. In this case inclusion probabilities are the same as using SπPS design, thus the notation of a such balanced sample is SπPS.

There are chosen four different sets of balanced variables for constructing the balance sample. These sets are notated by adding number to the name of balance sample:

-1. For this case there is one balance variable – inclusion probability;

-2. There are two balanced variables: inclusion probability and number of employees in the enterprise;

-3. There are two balanced variables: inclusion probability and tax of value added (VAT);

-4. There are three balanced variables: inclusion probability, number of employees in the enterprise and VAT.

For this research the three types of estimators are considered: Horvitz-Thompson (HT, see equation (2.29)), GREG-type (GREG, see equation (2.32)) and Model-based (MB, see equation (2.38)) estimators. For the GREG-type estimator, the predicted values are calculated using model $\mathcal{M}_9$ (see table 4.3) and for the MB estimator – $\mathcal{M}_{10}$ (see table 4.3). The models coefficients are estimated using the auxiliary information from the previous surveys. Thus, model coefficients are the same for all quarters, the auxiliary variables from the current survey are used just for estimation of the predicted values.

The results of this research are presented in tables 4.11, 4.12 and 4.13.

For the HT estimator the bigger effect has the difference between methods of the inclusion probabilities estimation than the number of balanced variables.

Table 4.11. Horvitz-Thompson estimates for different sample designs.

	Domain sample size classes					
Sample design	Small $0-9$		Medium $10-29$		Large $30-...$	
	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$
SRS-1	2.0	56.4	3.2	38.6	2.1	23.6
SRS-2	2.1	56.4	2.1	36.2	1.3	23.8
SRS-3	2.2	55.5	2.1	35.8	0.6	22.3
SRS-4	2.4	55.1	2.2	36.0	0.5	22.1
SSRS-1	2.4	31.6	0.7	15.2	0.4	8.8
SSRS-2	2.2	31.4	0.8	14.8	0.3	8.7
SSRS-3	1.7	30.5	0.8	15.1	0.4	8.0
SSRS-4	1.4	30.6	0.7	14.7	0.4	7.9
SπPS-1	2.3	52.4	0.7	18.8	0.4	8.4
SπPS-2	2.1	52.0	0.8	18.5	0.4	8.0
SπPS-3	1.6	51.1	0.7	18.4	0.3	7.9
SπPS-4	1.5	50.9	0.7	18.3	0.3	7.6

Table 4.12. GREG-type estimates under model $\mathcal{M}_9$ for different sample designs.

	Domain sample size classes					
Sample design	Small $0-9$		Medium $10-29$		Large $30-...$	
	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$
SRS-1	1.4	46.4	1.6	16.6	0.6	10.3
SRS-2	2.1	47.0	1.2	17.2	0.4	10.4
SRS-3	1.6	46.6	0.9	16.9	0.5	10.6
SRS-4	1.5	46.3	0.7	16.4	0.4	9.9
SSRS-1	0.8	20.0	0.7	12.7	0.2	5.5
SSRS-2	0.8	19.7	0.4	12.3	0.2	5.6
SSRS-3	0.7	19.4	0.4	12.3	0.2	5.4
SSRS-4	0.8	19.3	0.4	12.1	0.2	5.4
SπPS-1	0.8	34.0	0.8	17.0	0.5	7.4
SπPS-2	0.5	33.9	0.8	17.1	0.4	6.9
SπPS-3	0.5	33.2	0.7	16.6	0.3	6.7
SπPS-4	0.5	32.9	0.7	16.4	0.3	6.5

The same conclusions can be made and for GREG-type and MB estimators. For all estimators a balance sample with inclusion probabilities that are the same as for SSRS sample design presented best results. As for the estimators it seems that GREG-type estimator is better than MB estimator (especially for small domains). Still the previous research showed, that when nonresponse occurs MB estimator might have smaller MARB and RRMSE (see section 4.3.5). Thus the case when a balance sample is selected and nonresponse occurs is discussed in section 4.4.3.

Table 4.13. Model-based estimation under model $\mathcal{M}_{10}$ for different sample designs

	Domain sample size classes					
Sample design	Small $0-9$		Medium $10-29$		Large $30-\ldots$	
	$MABR, \%$	$MRRMSE, \%$	$MABR, \%$	$MRRMSE, \%$	$MABR, \%$	$MRRMSE, \%$
SRS-1	41.4	49.6	13.9	19.0	4.9	9.2
SRS-2	42.1	49.0	13.7	19.2	4.8	9.4
SRS-3	41.3	48.6	13.1	18.7	4.4	9.0
SRS-4	40.9	48.3	12.8	18.4	4.3	8.9
SSRS-1	19.2	23.3	7.5	14.0	2.2	5.1
SSRS-2	19.1	23.4	7.4	14.1	2.2	5.1
SSRS-3	18.5	22.7	7.2	13.5	2.2	5.0
SSRS-4	18.3	22.3	7.0	13.1	2.1	4.9
SπPS-1	23.8	27.9	10.2	14.7	3.1	6.0
SπPS-2	23.5	27.6	10.1	14.6	3.0	5.9
SπPS-3	23.6	27.7	10.0	14.2	3.3	6.0
SπPS-4	23.2	27.3	9.7	14.0	2.8	5.8

4.4.3. Balance sample and nonresponse

For comparison of the estimators when nonresponse occurs the a balance sample with inclusion probabilities that are the same as for SSRS sample design is used. A re-weighting method where response probabilities are estimated using logistic regression model is taken as the best nonresponse adjustment method for small area estimation (see section 4.3.5).

The results presented in table 4.14 show that the bias of GREG-type estimator increases quicker than model-based estimator when response rate increases.

Table 4.14. GREG-type and model-based estimators behavior when different response rate occurs

	Domain sample size classes					
Response rate	Small $0-9$		Medium $10-29$		Large $30-\ldots$	
	$MABR, \%$	$MRRMSE, \%$	$MABR, \%$	$MRRMSE, \%$	$MABR, \%$	$MRRMSE, \%$
	GREG-type estimator					
100%	0.8	19.3	0.4	12.1	0.2	5.4
80%	3.6	21.4	3.0	13.4	1.7	6.1
70%	6.7	25.2	4.6	16.0	2.2	7.2
	Model-based estimator					
100%	18.3	22.3	7.0	13.1	2.1	4.9
80%	18.5	22.9	7.2	13.2	2.2	5.1
70%	19.5	24.1	7.4	13.9	2.3	5.4

The same conclusion is made and in section 4.3.5, but this time, when a balance sample is used all the accuracy measures are smaller.

4.5. Model-based sample design and estimators

The results in section 4.4.2 show that accuracy measures depends more on the way the inclusion probabilities are constructed than on the number of balanced variables. Thus in this section model-based sample design is used to select sample. The first model-based sample design step says that an appropriative model should be selected and estimation of it's coefficients should be done. The coefficients should be estimated using auxiliary information which is available in the sample construction stage.

Six panel type models ($\mathcal{M}_5$–$\mathcal{M}_{10}$) are used in previous researches (see table 4.3), thus the same models are used and in this research.

After the model is selected and coefficients are estimated, the variance of the prediction error is calculated for all units in the population. This variable is used for the construction of inclusion probabilities, which are calculated in two different ways:

1. The population is divided into three strata by the value of the variance of the prediction error. The number of selected units from each strata is estimated using Neyman allocation formula (Särndal, Swensson, Wretman (1992)) where the variance is estimated not for a study variable but for the variance of the prediction error. The inclusion probabilities for the units from the same strata are equal. Thus using such sample design, the units from these strata where the variance of the prediction error is large have bigger probability to be selected, then the units from these strata where the the variance of the prediction error is small.

2. The population is divided into strata and number of elements from each strata are estimated in the same way as in the first case. Thus the inclusion probabilities in each strata are calculated proportional to the variance of the prediction error. Thus using such sample design, the units where the variance of the prediction error is large have bigger probability to be selected, then the units where the the variance of the prediction error is small.

Two estimators (GREG-type and model-based) are used to estimated parameter of interest. The model used in the estimation stage is the same as in the sample selection stage. The results are presented in tables 4.15 and 4.16.

Table 4.15. Results for model-based sample design when inclusion probabilities are the same for all units in the same strata.

Model	Domain sample size classes					
	Small $0-9$		Medium $10-29$		Large $30-\ldots$	
	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$
	GREG-type estimator					
$\mathcal{M}_5$	1.4	19.5	0.7	12.3	0.2	5.6
$\mathcal{M}_6$	1.3	19.2	0.6	12.1	0.2	5.4
$\mathcal{M}_7$	1.3	19.2	0.6	12.1	0.2	5.4
$\mathcal{M}_8$	1.0	19.0	0.5	12.0	0.1	5.2
$\mathcal{M}_9$	0.7	18.7	0.5	11.7	0.1	5.0
$\mathcal{M}_{10}$	0.8	18.7	0.5	11.8	0.1	5.0
	Model-based estimator					
$\mathcal{M}_5$	20.1	25.1	11.1	16.5	2.4	5.4
$\mathcal{M}_6$	18.2	23.3	9.5	15.0	2.3	5.2
$\mathcal{M}_7$	18.0	23.0	9.2	14.8	2.2	5.0
$\mathcal{M}_8$	16.0	21.8	7.0	13.1	2.1	4.9
$\mathcal{M}_9$	15.4	20.6	6.9	12.7	2.0	4.6
$\mathcal{M}_{10}$	15.2	20.3	6.6	12.5	2.0	4.4

Table 4.16. Results for model-based sample design when inclusion probabilities are proportional to the variance of prediction error.

Model	Domain sample size classes					
	Small $0-9$		Medium $10-29$		Large $30-\ldots$	
	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$	$MABR,\%$	$MRRMSE,\%$
	GREG-type estimator					
$\mathcal{M}_5$	1.5	19.4	0.6	12.2	0.2	5.5
$\mathcal{M}_6$	1.4	19.3	0.6	12.0	0.1	5.4
$\mathcal{M}_7$	1.4	19.3	0.5	12.0	0.1	5.4
$\mathcal{M}_8$	1.0	19.0	0.5	12.0	0.1	5.2
$\mathcal{M}_9$	0.8	18.8	0.4	11.6	0.1	5.0
$\mathcal{M}_{10}$	0.8	18.9	0.4	11.7	0.1	5.0
	Model-based estimator					
$\mathcal{M}_5$	19.3	23.1	10.3	14.6	2.2	5.2
$\mathcal{M}_6$	17.6	21.2	9.0	13.4	2.0	5.0
$\mathcal{M}_7$	17.4	20.4	8.8	13.3	1.9	4.8
$\mathcal{M}_8$	15.2	19.5	7.0	12.6	1.8	4.6
$\mathcal{M}_9$	14.6	18.7	6.5	11.8	1.5	4.1
$\mathcal{M}_{10}$	14.4	18.4	6.4	11.6	1.5	4.0

The comparison of the model-based sample designs demonstrates that for the GREG-type estimator there is no difference how the inclusion probabilities are constructed in the strata (in the table 4.15 they are the same and in the table 4.16 they are proportional to the variance of prediction error). Thus the difference between selected model is significant. For the GREG-type estimator panel type model with time component, $\mathcal{M}_9$, is the best.

The results also showed, that for model-based estimator the model-based sample design with inclusion probabilities proportional to the variance of prediction error (table 4.16) has smaller accuracy measures than model-based sample design with the same inclusion probabilities in the strata (table 4.16). The panel type model with time component and random intercept, $\mathcal{M}_{10}$, seems the best model for model-based estimator.

The analysis of sections 4.3.2, 4.4.2 and 4.5 reveals that model-based sample design for both estimators decrease accuracy measures more than other types of sample designs. Thus the optimal strategy for such kind of data is model-based sample design and GREG-type estimator.

4.6. The summary of the fourth chapter

The results of the first simulation showed, that:

1. Design-based estimators are approximately unbiased even when sample size is small. Thus the model-based estimators are biased and the bias might be so large, that relative root mean square error is larger then for the design-based estimators.

2. The use of panel type model for construction of superpopulation model and it's use for estimating small area estimators might improve the properties of estimator (a bias and a variance might be smaller) in comparison with the model constructed just using current's sample data.

3. The superpopulation models, which incorporate random effect improve the properties of model-based estimators in comparison with the models without random effects. For design-based estimator the improvement was not significant.

4. All estimators might be bias, if there is nonresponse. The bias depends on response rate: when response rate is small the bias is large. The bias grows quicker for design-based estimators in comparison with model-based estimators.

5. It is better not to use donor methods for nonresponse adjustments for small areas. It is better to use weighting methods or imputation using models.

The results of second simulation showed, that:

1. When the balance samples are used it is more important to properly choose inclusion probabilities, than the number of balancing variables. The results were better when balance sample with the same inclusion probabilities as SSRS design for all type estimators.

2. When the balance sample is used, the best estimation strategy is that where GREG estimator is used.

3. In the case of nonresponse, the properties of the estimators stay the same even when balance sample is used. Still estimators might be bias and it is better not to use donor method for nonresponse adjustment in small areas.

4. For model-based estimator the model-based sample design with inclusion probabilities proportional to the variance of prediction error has smaller accuracy measures than model-based sample design with the same inclusion probabilities in the strata. For design-based estimator both sample designs gave the same results.

5. For model-based sample design the best strategy is to use GREG estimator where panel type model without random effects and with time trend is used as the assisted tool. Still, if the model-based estimator is chosen, it is better to use panel type model with random effects and time trend as a superpopulation model.

General conclusions

After solving the problems formulated in the chapter "Introduction", we have obtained the following conclusions:

1. The analysis and simulation results using known sample designs and estimators showed that for small area estimation it is important to choose not only right estimator, but and the sample design and superpopulation model.

2. Not all nonresponse adjustment methods, which are used in population level, can be used and in small area level. For small area estimation it is suggested to use re-weighting or imputation using models, but not real donor imputation methods.

3. Simulation results showed, that the bias for design-based estimators increases quicker than for model-based estimators when response rate decreases.

4. Panel data model can be used to describe element's randomness and dependence on time in real finite population. Such type models are more useful than the models, which coefficients are estimated just using the current sample data.

5. Using balance sample it is more important to choose right inclusion probabilities than the right number of balanced variables. Simulation

results showed, that for all types of estimators it is better to use such balance sample, where inclusion probabilities are the same as stratified simple random sample, compared with other balanced samples.

6. The proposed model-based sample design might be the best choice, if a lot of information from previous researches are available. Simulation results showed, that the best estimation strategy might be where model-based sample design and GREG estimator are used. Here a panel data model with fixed-effects and time component should be used as assisted tool for GREG estimator. Still, if the model-based estimator is considered, the panel data model with random-effects and time component should be used.

References

Arora, V., Lahiri, P. 1988. On the superiority of the Bayesian method over the BLUP in small area estimation problems, *Statistics Sinica* 7: 1053–1063.

Battese, G. E., Harter, R. M., Fuller, W. A. 1988. An error component model for prediction of country crop areas using survey and satellite data, *Journal of the American Statistical Association* 83: 28–36.

Bethlehem, J. G. 1988. Reduction of nonresponse bias through regression estimation, *Journal of Official Statistics* 4: 251–260.

Brewer, K. R. W. 1999. Design-based or prediction-based inference? Stratified random vs stratified balanced sampling, *International Statistical Review* 67: 35–47.

Brick, J. M., Jones, M. E., Kalton, G. and Valliant, R. 2005. Variance Estimation with Hot Deck Imputation: A Simulation Study of Three Methods, *Survey Methodology* 31[2]: 151–159.

Brick, J. M., Kalton, G., Kim, J. K. 2004. Variance Estimation with Hot Deck Imputation Using a Model, *Survey Methodology* 30[1]: 57–66.

Chauvet, G. 2009. Stratified balanced sampling, *Survey Methodology* 35[1]: 115–119.

Chauvet, G., Tillé, Y. 2005. Fast SAS Macros for balancing Samples: user's guide, Software Manual, University of Neuchatel, http://www2.unine.ch/statistics/page10890.html.

Cumberland, W. G., Royall, R. M. 1981. Prediction models in unequal probability sampling, *Journal of the Royal Statistical Society* 43[B]: 353–367.

Cumberland, W. G., Royall, R. M. 1988. Does simple random sampling provide adequate balance?, *Journal of the Royal Statistical Society* 50[B]: 118–124.

Datta, G. S., Day, B. and Basawa, I. 1999. Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation, *Journal of statistical Planing and Inference* 75: 269–279.

Deville, J. C. 1992. Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information, In *Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys*, Orebro (Sweden).

Deville, J. C., Grosbras, J. M. and Roth, N. 1988. Efficient sampling algorithms and balanced sample, In COMPSTAT *Proceedings in Computational Statistics*, Heideberg, Physica Verlag, 255–266.

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

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.

Deville, J. C., Tillé, Y. 2005. Variance approximation under balanced sampling, *Journal of Statistical Planing and Inference*, 128: 569–591.

Ekholm, A., Laaksonen, S. 1991. Weighting via response modeling in the Finnish household budget survey, *Journal of Official Statistics*. 7: 325–337.

Eltinge, J. L., Yansaneh, I. S. 1997. Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey, *Survey Methodology* 23: 33–40.

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

Fay, R. E., Herriot, R. A. 1979. Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association* 74: 269–277.

Folsom, R. E., Singh, A. C. 2000. A generalized exponential model for sampling weight calibration for extreme values, non-response and poststratification, In *Proceedings of the Section on Survey Research Methods of the ASA* American Statistical Association, Alexandria, Virginia, 598–603.

Fuller, W. A. 2009. Some design properties of a rejective sampling procedure, *Biometrika*, 96: 933–944.

Fuller, W. A., An, A. B. 1998. Regression adjustment for nonresponse, *Journal of the Indian Society of Agricultural Statistics*, 51: 331–342.

Fuller, W. A., Harter, R. M. 1987. The multivariate components of the variance model for small area estimation, *Small Area Statistics* (Eds., R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh). New York: John Wiley and Sons, Inc., 103–123.

Ghosh, M., Rao, J. N. K. 1994. Small Area Estimation: An Appraisal, *Statistical Science* 9[1]: 55–93.

Groves, R., Dillman, D., Eltinge, J. and Little, R. 2002. *Survey Nonresponse*. Wiley, New York. 429 p.

Haziza, D., Thompson, K. J. and Yung, W. 2010. The effect of nonresponse adjustments on variance estimation, *Survey Methodology* 36: 35–43.

Hedayat, A. S., Majumdar, D. 1995. Generating desirable sampling plans by the technique of trade-off in experimental design, *Journal of Statistical Planning and Inference* 44: 237–247.

Horvitz, D. G.; Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* 47: 663–685.

Hsiao, C. 2003. *Analysis of Panel Data, Economic Society Monographs No. 34, 2nd Edition*. New York: Cambridge University Press. 369 p. ISSN 0-521-81855-9.

Iannacchione, V. G. 2003. Sequential weight adjustment for location and cooperation propensity for the 1995 national survey of family growth, *Journal of Official Statistics*, 19: 31–43.

Kalton, G. and Maligalig, D. S. 1991. A comparison of methods of weighting adjustment for nonresponse, *Proceedings of the U.S. Bureau of the Census Annual Research Conference*, 409–428.

Kim, J. K. and Kim, J. J. 2007. Nonresponse weighting adjustment using estimated response probability, *The Canadian Journal of Statistics*, 35[4]: 501–514.

Klaffe, J.; Rao, J. N. K. 1992. Estimation of mean square error of empirical best linear unbiased predictors under random error variance linear model, *Journal of Multivariate Analysis*, 43: 1–15.

Kott, P. S. 1986. When a mean-of-ratios is the best linear unbiased estimator under a model, *The American Statistician*, 40: 202–204.

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

Krapavickaitė, D.; Plikusas, A. 2005. *Imčių teorijos pagrindai*. Vilnius : Technika. 311 p. ISSN 9986-05-927-5.

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.

Lehtonen, R.; Särndal, C. E.; Veijanen, A. 2005. Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains, *Statistics in Transition* 7: 649–673.

Lehtonen, R.; Veijanen, A. 2009. Chapter 31 - Design-based Methods of Estimation for Domains and Small Areas, *Handbook of Statistics Sample Surveys: Inference and Analysis* 29B: 219–249.

Little, R. J. A. 1986. survey nonresponse adjustments for estimates of means, *International Statistical Review* 54: 139–157.

Lundström, S., Särndal, C. E. 1999. Calibration as a standart method for treatment of nonresponse, *Journal of Official Statistics* 15: 305–327.

Malec, D., Davis, W. W., Cao, X. 1999. Model-based small area estimates of overweight prevalence using sample selection adjustment, *Statistics in Medicine* 18: 3189–3200.

Malec, D., Sedransk, J., Moriarity, C. L. and Leclere, F. 1997. Small area inference for binary variables in the National Health Interview Survey, *Journal of the American Statistical association* 92: 815–826.

Moura, F., Holt, D. 1999. Small area estimation using multi level models, *Survey Methodology* 25: 73–80.

National Center for Heath Statistics. 1968. Synthetic State Estimates of Disability, PHS publication no. 1959. Washington, DC: Public Health Service, US Government Printing Office.

Narain, R. D. 1951. On sampling without replacement with varying probabilities, *Journal of the Indian Society of Agricultural Statistics* 3: 169–174.

Nedyalkova, D.; Tille, Y. 2008. Optimal Sampling and Estimation Strategies under The Linear Model, *Biometrika* 95[3]: 521–537.

Office of Management and Budget. Statistical Policy Office. Federal Committee on Statistical Methodology. Subcommittee on Small Area Estimation. 1993. *Indirect Estimators in Federal Programs,* Statistical Policy Working Paper 21. Washington.

Oh, H. L. and Scheuren, F. S 1983. Weighting adjustments for unit nonresponse, In *Incomplete Data in sample Surveys* Vol. 2, W.G. Madow, I. Olkin and D.B. rubin, eds. New York: Academic Press.

Omrani, H.; Gerber, P. and Bousch. P. 2009. Model-Based Small Area Estimation With Application To Unemployment Estimates, *World Academy Of Science, Engineering And Technology* 49: 793–800.

Rao, J. N. K. 1994. Estimating totals and distribution functions using auxiliary information at the estimation stage, *Journal of Official Statistics* 10[2]: 153–165.

Rao, J. N. K. 1999. Some recent advances in model-based small area estimation, *Survey Methodology* 25[2]: 175–186.

Rao, J. N. K. 2003. *Small Area Estimation.* Wiley, New York. 313 p. ISBN 0-471-41374-7.

Rao, J. N. K. 2005. Inferential Issues In Small Area Estimation: Some New Developments, *Statistics in Transition* 7: 513–526.

Rao, J. N. K. and Choudhry, G. H. 1995. Small area estimation: Overview and empirical study, *Business Survey Methods* (Eds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. christianson, M.J. Colledge, and P.S. Kott). New York: John Wiley and Sons, Inc., 527–542.

Rao, J. N. K. and Shao, J. 1992. Jackknife variance estimation with survey data under hot deck imputation, *Biometrika*, 79: 811–822.

Rao, J. N. K. and You, Y. 2002. A Pseudo-Empirical Best Hierarchical Bayes small area estimation using sampling weights, *Proceedings of the Survey Methods Section, Statistical Society of Canada,* 7: 513–526.

Royall, R. M. 1970. On Finite Population Sampling Theory under Certain Linear Regression Models, *Biometrika* 57[2]: 377–387.

Royall, R. M. 1976a. Likelihood functions in finite population sampling theory, *Biometrika* 63: 605–614.

Royall, R. M. 1976b. The linear least squares prediction approach to two-stage sampling, *Journal of the American Statistical Association* 71: 657–664.

Royall, R. M. 1988. The prediction approach to sampling theory, In *Handbook of Statistics Volume 6: Sampling* (Eds., P.R. Krishnaiah and C.R. Rao), Amsterdam. Elsevier/North-Holland, 399–413.

Royall, R. M. and Pfeffermann, D. 1982. Balanced samples and robust bayesian inference in finite population sampling, *Biometrika* 69: 401–409.

Rosenbaum, P. R. 1987. Model-based direct adjustment, *Journal of the american Statistical Association* 89: 846–866.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons. 320 p.

Saei, A.; Chambers, R. 2003. Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects, *Methodology Working Paper No. M03/15.*

Särndal, C. E. 1996. Efficient estimators with simple variance in unequal probability sampling, *Journal of the American Statistical Association* 91: 1289–1300.

Särndal, C. E. 2007. The calibration approach in survey theory and practice, *Survey Methodology* 33[2]: 99–119.

Särndal, C. E. 2007. Topics in uses of auxiliary information in surveys: The role of models, Nonresponse adjustment, Estimation for (small) domains, *Proceedings, Second Baltic-Nordic Conference on Survey Sampling*.

Särndal, C. E.; Lundstrom, S. 2005. *Estimation in Surveys with Nonresponse*. John Wiley and Sons. 212 p. ISBN 978-0-470-01133-1.

Särndal, C. E.; Swensson, B.; Wretman, J. 1992. *Model Assisted Survey Sampling*. Springer-Verlag. 694 p. ISBN 0-387-97528-4.

Schaible, W. L. 1992. Use of small area statistics in U.S. Federal Programs, In *Small area Statistics and Survey Designs* (G. Kalton, J. Kordos and R. Platek, eds.) 1: 95–114. Central Statistical Office, Warsaw.

Singh, M. P.; Gambino, J.; Mantel, H. J. 1994. Issues and strategies for small area data, *Survey Methodology* 20: 3–14.

Singh, M. P.; Stukel, D. M.; Pfeffermann, D. 1998. Bayesian versus frequentist measures of error in small area estimation, *Journal of the Royal statistical Society*, Series B 60: 377–396.

Tillé, Y. 2011. Ten years of balanced sampling with the cube method: An appraisal, *Survey Methodology* 37[2]: 215–226.

Tillé, Y.; Matei, A. 2007. *The R Package Sampling*. The Comprehensive R Archive Network, http://cran.r-project.org/, Manual of the Contributed Packages.

Stukel, D. M.; Rao, J. N. K. 1999. On small-area estimation under two-fold nested error regression models, *Journal of Statistical Planing and Inference*, 78: 131–147.

U.S. Office Of Management And Budget. 1993. Indirect Estimators in Federal Programs. U.S. Office of Management and Budget, Statistical Policy Working Paper 21, National Technical Information Service, Springfield, Virginia.

Yates, F. 1946. A review of recent statistical developments in sampling and sampling surveys, *Journal of Royal Statistical Society*, A109: 12–43.

List of scientific publications on the topic of the dissertation

In the reviewed scientific journals

Nekrašaitė-Liegė, V. 2011a. Some applications of panel data models in small area estimation, *Statistics in transition – new series*, 12(2): 265–280, ISSN 1234-7655.

Nekrašaitė-Liegė, V.; Radavičius, M.; Rudys, T. 2011b. Model-based design in small area estimation, *Lithuanian Mathematical Journal*, 51(3): 417–424. ISSN 0363-1672. (Thomson Reuters Web of Knowledge).

In other scientific journals

Nekrašaitė-Liegė, V. 2012. Estimation strategy for small areas, a case study, In *Proceedings of the Workshop on Survey Sampling Theory and Methodology, held in Valmiera on 24–28 August, 2012*, p. 143–147. ISBN 978-9984-45-557-0.

Nekrašaitė-Liegė, V. 2010a. Weighting and imputation comparison in small area estimation. *Lietuvos matematikos rinkinys. LMD darbai*, 51:414–419. ISSN 0132-2818.

Nekrašaitė-Liegė, V. 2010b. Nonresponse adjustment in SAE under different sampling designs. In *Proceedings of the Workshop on Survey*

Sampling Theory and Methodology, held in Vilnius on 23–27 August, 2010. Vilnius: Statistics Lithuania, p. 137–141. ISBN 978-9955-797-06-7.

Nekrašaitė-Liegė, V. 2009. Small area estimation in the case of non-response. *Lietuvos matematikos rinkinys. LMD darbai*, 50:304–309. ISSN 0132-2818.

Vilma Nekrašaitė-Liegė

SMALL AREA ESTIMATION
Doctoral Dissertation
Physical Sciences, Mathematics (01P)

MAŽŲ SRIČIŲ VERTINIMAS,
Daktaro disertacija
Fiziniai mokslai, matematika (01P)

2012 12 21. 9,0 sp. l. Tiražas 20 egz.
Vilniaus Gedimino technikos universiteto leidykla „Technika",
Saulėtekio al. 11, 10223 Vilnius, *http://leidykla.vgtu.lt*
Spausdino UAB „Ciklonas",
J. Jasinskio g. 15, LT-01111 Vilnius