CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Nuclear Sciences and Physical Engineering

HABILITATION THESIS

# Model-based methods for small area estimation

Tomáš Hobza

Prague 2017

# ABSTRAKT V ČESKÉM JAZYCE


**Odhadování v malých oblastech založené na statistických modelech**

Tato práce shrnuje hlavní výsledky autora získané v tematice zvané „odhadování v malých oblastech“ (anglicky „small area estimation“ - SAE). SAE je odvětví matematické statistiky, které se zabývá odhadováním parametrů v podmnožinách (nazývaných oblasti nebo domény) jisté populace, ve kterých není k dispozici dostatečné množství dat pro spolehlivé přímé odhady. Za tímto účelem zavádí SAE modely, které si „půjčují sílu“ ze souvisejících malých oblastí, z dat získaných z externích administrativních zdrojů nebo z dat z jiných časových období. Přehled základních principů, modelů a problémů vyskytujících se v SAE je obsahem první části práce.

Příspěvek autora k této problematice je ve formě publikovaných článků prezentován ve druhé části práce a spočívá v několika modelech navržených pro odhadování parametrů malých oblastí. Tyto modely jsou založeny na lineárních smíšených a zobecněných lineárních smíšených modelech. Konkrétně je navrženo a studováno několik modifikací Fay-Herriotova modelu a regresního modelu se vnořenými chybami a dále jsou uvažovány logistické smíšené modely pro binární data. Pro všechny uvažované modely jsou řešeny následující problémy. Jsou odvozeny formule a algoritmy pro odhadování neznámých parametrů modelu. Jsou studovány empirické nejlepší prediktory parametrů malých oblastí založené na studovaných modelech. Zvláštní pozornost je věnována odhadům střední kvadratické chyby prediktorů, neboť takováto míra přesnosti je potřebná pro praktické aplikace. Pro všechny modely jsou popsány analytické aproximace středních kvadratických chyb nebo jejich odhady pomocí metody bootstrap.

Důležitou součástí vytváření nového modelu je návrh a realizace simulačních experimentů studujících chování nových metod pro malé rozsahy výběrů a porovnávající je s již existujícími metodami, pokud nějaké jsou. Nedílnou součástí prezentovaných prací je tedy vývoj netriviálních softwarových nástrojů, neboť standardní statistické balíky nelze pro studované modely použít a navíc je třeba často použít aproximace pomocí metody Monte Carlo. Aby byla ukázána aplikovatelnost a přínos navržených metod v praxi, je ve všech článcích provedena aplikace na reálná data.

Závěrem je uvedeno několik výsledků autora, týkajících se robustního odhadování a detekce odlehlých pozorování v zobecněných lineárních modelech, které mohou být aplikovány na problémy odhadování v malých oblastech.

# ABSTRACT IN ENGLISH


## Model-based methods for small area estimation

This thesis summarizes main results of the author obtained in the field called "small area estimation" (SAE). SAE is a branch of mathematical statistics which deals with the problem of estimating population parameters in subsets (called areas or domains) of a population where the sample sizes are not large enough to provide reliable direct estimates. For this purpose, SAE introduces statistical models that "borrow strength" from related small areas, data from external administrative sources or data from different time periods. An overview of basic principles, models and problems encountered in SAE is given in the first part of the thesis.

The contribution of the author in the form of published papers is presented in the second part of the thesis and consists of several models proposed for estimation of small area parameters. These models are based on linear mixed and generalized linear mixed models. Namely, there are proposed and studied several modifications of Fay-Herriot and nested error regression models and there are considered logistic mixed models for binary data. For all the assumed models the following problems are treated. Formulas and algorithms for estimation of the unknown parameters of the models are derived. Model-based empirical best predictors of parameters of small areas are studied. Special attention is paid to estimation of the mean squared error of the predictors since such a measure of accuracy is needed in practical applications. For all the models analytic approximation or bootstrap estimates of the mean squared errors are given.

An important part of developing a new model is to design and carry out simulation experiments studying small sample size behaviour of the new methods and comparing them with the existing ones if there are any available. It means that development of non-trivial software tools is an integral part of the presented works since the standard statistical packages cannot be used for the studied models and moreover Monte Carlo approximation methods must often be used. Further, to show applicability and benefits of the proposed methods in practice, a real data application are performed in all the papers.

In addition, several results of the author which are connected with robust estimation and outlier detection in generalized linear models and which can be applied to small area estimation problems are presented.

# Contents

**Introduction** **7**

**1 Definition of the problem and basic concepts** **9**
1.1 Design-based methods . . . . . 10
1.2 Model-based small area estimation . . . . . 14

**2 Linear mixed models** **17**
2.1 Estimation of parameters of linear mixed models . . . . . 17
2.1.1 MLE . . . . . 18
2.1.2 REML . . . . . 18
2.2 Prediction of linear combination of model effects . . . . . 19
2.3 Prediction of linear combination of observations $\boldsymbol{y}$ . . . . . 19
2.4 Mean squared error of EBLUP . . . . . 20

**3 Models for continuous responses** **23**
3.1 Area level models . . . . . 23
3.1.1 Empirical best predictors . . . . . 24
3.1.2 MSE of EBLUP . . . . . 24
3.1.3 Contribution of the author to area level models . . . . . 25
3.2 Unit level models . . . . . 27
3.2.1 Empirical best predictors . . . . . 27
3.2.2 Mean squared error of EBLUP . . . . . 29
3.2.3 Contribution of the author to unit level models . . . . . 30

**4 Models for discrete responses** **33**
4.1 Generalized linear mixed models . . . . . 33
4.2 Unit level logistic mixed model . . . . . 34
4.2.1 Estimation of parameters of logistic mixed model . . . . . 35
4.2.2 Prediction of functions of fixed and random effects . . . . . 36
4.2.3 Mean squared error of predictors . . . . . 38
4.3 Contribution of the author to unit level logistic mixed models . . . . . 39
4.4 Contribution of the author to unit level generalized linear models . . . . . 41

**Conclusion and possible future directions** **45**

**References** **47**

**5 Relevant published articles of the author** **51**
5.1 A Fay-Herriot model with different random effect variances . . . . . 51
5.2 An Area-Level Model with Fixed or Random Domain Effects in Small Area Estimation Problems . . . . . 65
5.3 A modified nested-error regression model for small area estimation . . . . . 78
5.4 Small Area Estimation of Poverty Proportions under Random Regression Coefficient Models . . . . . 95
5.5 Small area estimation under random regression coefficient models . . . . . 110
5.6 Empirical best prediction under unit-level logit mixed models . . . . . 129
5.7 Robust Median Estimator in Logistic Regression . . . . . 162
5.8 Robust median estimator for generalized linear models with binary responses . . 182
5.9 Outlier detection method in GEEs . . . . . 210

**Appendices**

**A Curriculum vitae** **227**

**B List of publications** **231**

# Introduction

Survey-sampling is widely used in practice for obtaining information on a wide range of topics of interest and its methodology is developing rapidly with increasing value of quantitative data and major developments in computing power. At the beginning it was used to provide estimates mainly for the total population under consideration (e.g. all inhabitants of a certain state), but over time the demand for reliable estimates for a variety of subpopulation (domains) has appeared and greatly increased. Domains may represent some geographic areas such as states, provinces, counties, municipalities etc., or socio-demographic groups such as a specific age-sex-race groups of people in a large geographic area. In this context, an estimator of a domain parameter is called *"direct estimator"* if it is calculated just with the sample or auxiliary data coming from the corresponding domain.

The overall sample size of a sample survey is usually determined to provide specific accuracy of direct estimates for large geographical regions or broad demographic groups. But due to budget restrictions and other constraints it is often not possible to have enough data to support reliable direct estimators for all considered subpopulations (e.g. counties). Similar problems may arise when all potential uses of the survey data are not anticipated at the design stage of the study and new requirements or the level of domains of interest are specified after finishing the survey, when it is almost impossible or very expensive to repeat the data collection process and get the additional information. So in practical applications we often find situations, where many domains of interest have very small or even zero sample size. Such domains, not having large enough sample sizes for producing direct estimates of adequate precision, are called *"small areas"*.

*Small area estimation* (SAE) is thus a field of mathematical statistics dealing with the problem of obtaining reliable estimates of characteristics of interest (means, totals, quantiles etc.) for domains for which only small samples or no samples are available, i.e. for small areas. Of course it is not sufficient to provide just point estimates of some characteristic. A second problem is how to asses the estimation error. In order to obtain estimates for small areas the idea is to "borrow strength" by using variables from related or similar small areas and to formulate "indirect" estimators that increase the effective sample size. Indirect estimators are based on a model that provides a link to related small areas through auxiliary data obtained from external sources such as large surveys, recent census or current administrative records.

SAE methods can be generally divided into "design-based" and "model-based" methods. The design-based methods make use of survey-weights and they often employ an implicit model for the construction of the estimators. However, the bias, the variance and other properties of the estimators are evaluated under the probability distribution induced by the sampling design used to select the sample. Under this setup, the population values are supposed to be fixed. On the other hand, the model-based methods use an explicit model, treat the population values as random, usually condition to the selected sample, and the inference properties of the estimators are optimized with respect to the underlying model distribution. The latter methods use either the frequentist approach or the Bayesian methodology. A common crucial feature to both SAE approaches is the availability of good auxiliary data. Without having a set of covariates with

a good predictive power for the small area parameters of interest, even the most complex and elaborated models can be of little help with the small samples often encountered in practice.

Although the history of SAE goes back to the eleventh century England or seventeenth century Canada as reported in Brackstone (1987), these early small area statistics were all based on administrative records aiming at complete enumeration, and the real development of SAE methodology can be dated to the last decade of the 20th century. Since then, the SAE is flourishing both in research and applications. This is due to the demand for reliable small area statistics coming from both the public and private sectors, which has been increasing worldwide in recent years. The information obtained by SAE methods is used in regional and urban planning, allocation of funds in many government programs covering education, public health, poverty etc. In developing countries e.g. there is an increasing demand in governmental agencies for income and poverty estimators in small areas. Another field of application is the so called disease mapping when small area techniques are used to predict disease incidence over different small areas which can help to identify factors (such as environmental pollution) causing a disease. The importance of small area statistics has increased significantly also in the private sector since many business decisions rely on the local socio-economic conditions. Of course this rapid development of SAE methods is also connected with the increasing amount and quality of collected data and with the advances in statistical data processing. Nowadays, high-speed computers allow fast processing of large data sets and applications of much more complex and complicated models than in the past.

The early development of SAE was done mainly in the USA (and Canada) where the Census Bureau formed a committee on Small Area Income and Poverty Estimates in the early's 1990 with the aim of providing estimates of income and poverty at state, county, and district levels. Another examples are the Local Area Unemployment Statistics program of the Bureau of Labor Statistics producing monthly estimates of unemployment rates for states, metropolitan areas, and counties or the County Estimates Program of the National Agricultural Statistics Service producing county estimates of crop yield. The European Union also did not want to stay away from this increasing trend of using small area statistics and several European research projects such us EURAREA, SAMPLE and AMELI have been supported by the European Commission. Actually, the small area methods are used in the frame of the program "European Statistics on Income and Living Conditions" (EU-SILC) which is one of the statistical operations that have been harmonised for EU countries. The main goal of the Living Conditions Surveys (LCS) is to provide a reference source on comparative statistics on the distribution of income and social exclusion in the European environment.

More information about the small area estimation and its main developments during the recent years can be found in the monographs of Rao (2003) and Rao and Molina (2015), and the reviews of Ghosh and Rao (1994), Rao (1999), Pfeffermann (2002, 2013), and Jiang and Lahiri (2006).

This thesis deals mainly with model-based methods under the frequentist approach and is organized as follows. Chapter 1 introduces the small area estimation problem and two main approaches to its solution. Chapter 2 presents basic principles of linear mixed models which are then used in Chapter 3 for description of two types of small area models for continuous responses. Chapter 4 deals with small area models for discrete responses based on generalized linear mixed models. The main purpose of this initial part of the thesis is to give a brief overview of basic problems, techniques and models which are encountered in small area estimation. At the same time the contribution of the author to each of the presented type of models is shortly explained. Chapter 5 presents relevant articles of the author, namely seven published impacted papers and two papers published in a book. The thesis contains also two appendices which give the CV and list of publications of the author.

# Chapter 1

# Definition of the problem and basic concepts

Consider a finite population $U$ of size $N$ which is partitioned into $D$ domains or areas denoted as $U_d$ with $N_d$ units in area $d$, so that $U = \cup_{d=1}^{D} U_d$ and $N = \sum_{d=1}^{D} N_d$. By population we mean a collection of distinct units like persons, households, companies, hospitals etc. which can be identified through the labels $j = 1, \ldots, N$. An area may represent a concrete geographical area or a socio-economic group. Let $y$ denote the characteristic of interest (e.g. personal income, indicator if a person is unemployed etc.) and $y_{dj}$ the corresponding value of characteristic $y$ for the $j$-th unit in area $d$, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$. Under this notation, in each area $d$ there exists the vector

$$\boldsymbol{y}_d = (y_{d1}, \ldots, y_{dN_d})^T, \qquad d = 1, \ldots, D\,,$$

containing the values of $y$ associated with the units of area $d$.

From the population a sample $s \subset U$ of size $n$ is selected. Let $s = s_1 \cup \ldots \cup s_D$, where $s_d$, $d = 1, \ldots, D$, defines the sample observed for area $d$ with corresponding sample sizes $n_d$ satisfying $n = \sum_{d=1}^{D} n_d$. The sample sizes $n_d$ may be generally random unless a planned sample of fixed size is taken in each area. Without loss of generality, we assume that the sample in area $d$ consists of the first $n_d$ units of the subpopulation $U_d$ so that the vector $\boldsymbol{y}_d$ can be written in the form $\boldsymbol{y}_d = (\boldsymbol{y}_{ds}^T, \boldsymbol{y}_{dr}^T)^T$, where

- $\boldsymbol{y}_{ds} = (y_{d1}, \ldots, y_{dn_d})^T$ is the vector corresponding to the $n_d$ observed units in area $d$
- $\boldsymbol{y}_{dr}$ is the vector corresponding to the $N_d - n_d$ unobserved units in area $d$.

Now we are ready to state the basic task of small area estimation which is twofold.

**General problem:**

1. How to estimate, on the basis of the selected sample $s$, for each area $d$ the quantity

$$h(y_{d1}, \ldots, y_{dN_d})\,,$$

where $h$ is a known function.

2. How to express uncertainty of this estimate.

The function $h$ may be linear in which case the two most typical examples of the target

quantity are the *population total* in area $d$

$$Y_{d\cdot} \triangleq \sum_{j=1}^{N_d} y_{dj}, \qquad d = 1, \ldots, D\,,$$

and the *population mean* in area $d$

$$\overline{Y}_d \triangleq \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}, \qquad d = 1, \ldots, D\,.$$

E.g. we may be interested in estimating the totals of unemployed people for each area or the mean personal income for each area.

In order to give an example of used nonlinear function $h$ let us mention for example the family of so called FGT poverty measures which are defined for each area $d$ as the area mean

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} F_{\alpha dj}, \qquad d = 1, \ldots, D\,,$$

of the values $F_{\alpha dj}$ defined as

$$F_{\alpha dj} = \left(\frac{z - y_{dj}}{z}\right)^{\alpha} I(y_{dj} < z), \qquad j = 1, \ldots, N_d,\ \alpha = 0, 1, 2,$$

where $I(y_{dj} < z)$ is the indicator function taking on value 1 if $y_{dj} < z$ and value 0 otherwise. To explain the meaning of this family of measures imagine that $y_{dj}$ is some measure of welfare for individual $j$ in area $d$, such as income or expenditure, and $z$ is a given fixed poverty line. It means that $z$ is the threshold for $y_{dj}$ under which a person is considered "to be poor". E.g. EUROSTAT defines the poverty line as the 60% of the median of the equivalent personal income in the whole country or region during the previous year. Under this setup, for $\alpha = 0$ the measure expresses the proportion of individuals under poverty line in area $d$ and it is called *poverty incidence*. For $\alpha = 1$ the measure is called *poverty gap*, and calculates the area mean of the relative distance to the poverty line of each individual. The FGT measure for $\alpha = 2$ is called *poverty severity* and its large values should point out to areas with severe level of poverty.

In small area estimation there are basically two approaches to estimation of the target characteristics of areas on the basis of the selected sample, namely the design-based approach and the model-based approach. Since this thesis deals mainly with the latter one, we mention now just basic principles of the traditional design-based approach.

## 1.1 Design-based methods

The basic feature of the design-based methods is that they suppose the population values $y_{dj}$, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$, to be fixed and the randomness is incorporated by the random selection of the sample $s$. In practice, a probability *sampling design* is used to select a sample. It is in fact a scheme for choosing the sample so that every subset $s$ of the population $U$ has a known probability $p(s)$ of selection. In general, it is difficult to calculate the probability $p(s)$. Some simple cases are

$$p(s) = \frac{1}{\binom{N}{n}} \qquad \text{or} \qquad p(s) = \frac{1}{N^n} \tag{1.1}$$

for a sample of size $n$ under simple random sampling without replacement or simple random sampling with replacement, respectively. Of course, usually more complicated designs are used as e.g. stratified simple random sampling or stratified multistage sampling.

All inferences under the design-based approach are done with respect to the selection probabilities $p(s)$ (sometimes called *randomization distribution*) of the sample $s$. For instance, the definition of bias of some estimator $\widehat{T}$ of a quantity $T$, which is evaluated on the basis of sample $s$, is

$$\mathrm{E}_D(\widehat{T} - T) = \sum_{s \subset U} p(s) \left( \widehat{T}(s) - T \right) ,$$

where the summation is over all possible samples $s$ that can be drawn from the population using a particular sampling design. The subscript $D$ indicates that the expectation is taken with respect to a sampling design and it is not based on a model as in the next section. Similarly, the variance of an estimator $\widehat{T}$ is defined as

$$\mathrm{V}_D(\widehat{T}) = \sum_{s \subset U} p(s) \left( \widehat{T}(s) - \mathrm{E}_D(\widehat{T}) \right)^2 .$$

In order to derive theoretical properties of an estimator it is usually necessary to work with the so called *inclusion probabilities* $\pi_j$ of individual $j$ rather than with the selection probabilities $p(s)$ of the sample $s$. Let us explain their meaning. The probability $\pi_j$ that a unit $j$, $j = 1, \ldots, N$, will be in the selected sample $s$ is

$$\pi_j = \sum_{s \in s(j)} p(s) ,$$

where $s(j)$ stand for the set of all potential samples that contain unit $j$, i.e. $s(j) = \{s \subset U \,|\, j \in s\}$. For calculating the sampling variance the joint inclusion probability that a unit $i$ and unit $j$ will be in the selected sample is needed and it can be expressed as

$$\pi_{i,j} = \sum_{s \in s(i,j)} p(s) ,$$

where $s(i,j) = \{s \subset U \,|\, i, j \in s\}$. In sample surveys the so called "sampling weights" $w_j$ play an important role in constructing design-based estimators. An important basic choice of the weight for individual $j$ is $w_j = 1/\pi_j$, where $\pi_j$ is the inclusion probability of the individual. The weight $w_j$ may be interpreted as the number of elements in the population represented by the sample element $j$.

**Example 1** One of the simplest sampling designs is the simple random sampling without replacement. It assumes that no unit can appear in the sample more than once and assigns the same probabilities of selection $1/\binom{N}{n}$ (cf. (1.1)) to each of the $\binom{N}{n}$ sets of $n$ different units from the population of size $N$. In this case, to calculate the inclusion probability for unit $j$ it suffices to evaluate the number of potential samples that contain unit $j$, i.e. the size of the set $s(j)$. But given that unit $j$ is in the sample, the rest of $n-1$ sample units must be selected from the remaining $N-1$ units in the population and this can be done in $\binom{N-1}{n-1}$ different ways. Thus

$$\pi_j = \sum_{s \in s(j)} p(s) = \sum_{s \in s(j)} \frac{1}{\binom{N}{n}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

and the corresponding sampling weight for unit $j$ is $w_j = N/n$. Let us note that in a similar way we may obtain the joint inclusion probability of units $i$ and $j$ in the form

$$\pi_{i,j} = \frac{n}{N} \frac{n-1}{N-1} .$$

Let us now return to the general problem of estimating some function $h(y_{d1}, \ldots, y_{dN_d})$ on the basis of sample $s$. We split the discussion into two parts with respect to the used sampling design.

**I)** For simplicity, let us first suppose that within each domain $d$ a sample of size $n_d$ is selected by simple random sampling without replacement and that the target quantities of interest are the means $\overline{Y}_d$. If there is no additional information available, the *direct estimator* of the area mean $\overline{Y}_d$ is

$$\widehat{\overline{Y}}_d^{dir} = \overline{y}_d = \frac{1}{n_d} \sum_{j=1}^{n_d} y_{dj}, \qquad d = 1, \ldots, D, \tag{1.2}$$

and its design variance over the randomization distribution is given by

$$\mathrm{V}_D\left(\overline{y}_d\right) = \frac{S_d^2}{n_d}\left(1 - \frac{n_d}{N_d}\right), \quad \text{where} \quad S_d^2 = \frac{1}{N_d - 1}\sum_{j=1}^{N_d}(y_{dj} - \overline{Y}_d)^2\,. \tag{1.3}$$

The term "direct" is used to express that the estimator uses only the sample data coming from the target area. The direct estimator is design-unbiased, i.e.

$$\mathrm{E}_D\left(\widehat{\overline{Y}}_d^{dir}\right) = \overline{Y}_d\,,$$

but from the formula of the design variance one can see that for small sample sizes $n_d$ the variance will be large, unless the variability of the $y$-values, $S_d^2$, is sufficiently small. This is the point where the problem of small area estimation arises. If there are domains where the sample sizes $n_d$ are small in our sample, the direct estimates in these areas will not have adequate precision because its variance will be large.

One possibility how to decrease the variability of direct estimators is to use some additional auxiliary data. Suppose that vector of covariates $\boldsymbol{x}_{dj} = (x_{dj,1}, \ldots, x_{dj,p})^T$, $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, is also known for each unit in the sample $s$ and that the population area means $\overline{\boldsymbol{X}}_d = 1/N_d \sum_{j=1}^{N_d} \boldsymbol{x}_{dj}$ are known as well. This additional information may be obtained for example from recent census or some other administrative registers. In such case more efficient estimator called *synthetic regression* estimator can be defined as

$$\widehat{\overline{Y}}_d^{sr} = \overline{y}_d + \left(\overline{\boldsymbol{X}}_d - \overline{\boldsymbol{x}}_d\right)^T \widehat{\boldsymbol{\beta}}\,,$$

where $\overline{\boldsymbol{x}}_d = 1/n_d \sum_{j=1}^{n_d} \boldsymbol{x}_{dj}$ are the sample means of the covariates and

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D}\sum_{j=1}^{n_d} \boldsymbol{x}_{dj}\boldsymbol{x}_{dj}^T\right)^{-1}\left(\sum_{d=1}^{D}\sum_{j=1}^{n_d} \boldsymbol{x}_{dj} y_{dj}\right)$$

is the ordinary least square estimator of unknown parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. This estimator is motivated by a linear regression model of the $y$ values on covariates $x$ in the population with a common vector of regression coefficients, it thus assumes that the areas are homogeneous with respect to the estimated quantity, i.e. the vector of regression parameters $\boldsymbol{\beta}$ is similar across the areas. The term "synthetic" refers to the fact that an estimator $\widehat{\boldsymbol{\beta}}$ computed from data of all the areas is used for every area separately, borrowing thus information from other "similar areas". Therefore the synthetic estimators are indirect estimators and they are sometimes called *"model-assisted"*.

The main advantage of the synthetic regression estimator over the direct estimator is that its design-based variance $\mathrm{V}_D\left(\widehat{\overline{Y}}_d^{sr}\right)$ is of the order $O(1/n)$, where $n = \sum_{d=1}^{D} n_d$ is usually large,

while the design-based variance of the direct estimator is of the order $O(1/n_d)$ (cf. (1.3)) and $n_d$ may be small. Hence, the synthetic regression estimator may decrease the variance substantially. On the other hand, however, it may lead to a large bias if the assumption of homogeneity of the regression coefficients within the population is not fulfilled.

To make compromise between the direct estimator with small or no bias but large variance and the synthetic regression estimator with small variance but possibly large bias, sometimes a linear combination of the two is assumed. The resulting estimator

$$\widehat{\overline{Y}}_d^{com} = \tau_d \widehat{\overline{Y}}_d^{dir} + (1-\tau_d)\widehat{\overline{Y}}_d^{sr}, \qquad 0 \le \tau_d \le 1,$$

is called *composite* estimator. Here the main question is the choice of the weights $\tau_d$ which should be ideally selected so that the mean square error (MSE) of the resulting estimator is minimized. But this is problematic since it is difficult to derive the bias of the synthetic estimator accurately and the MSE is thus generally unknown. So the usual choice of the weights $\tau_d$ typically depends on the sample size $n_d$ of the area $d$ in such a way that in areas with larger sample size $n_d$ more weight is put to the direct estimator. For more details on specifying the weights and other types of composite estimators see Rao and Molina (2015).

**II)** If the sampling design is more complex and sampling weights $w_{dj}$ for the $j$-th unit in area $d$ are used, the direct estimator of the mean $\overline{Y}_d$ is defined as

$$\widehat{\overline{Y}}_d^{dir} = \frac{1}{\widehat{N}_d} \sum_{j=1}^{n_d} w_{dj} y_{dj},$$

where $\widehat{N}_d = \sum_{j=1}^{n_d} w_{dj}$. Note that for the choice $w_{dj} = N_d/n_d$, i.e. when the sample is selected by simple random sampling without replacement within each area $d$, this estimator corresponds to the ordinary direct estimator defined in (1.2). Since this estimator is used for practical applications in some of the attached papers we also give formula for estimating its design-based variance, namely

$$\widehat{\mathrm{V}}_D\left(\widehat{\overline{Y}}_d^{dir}\right) = \frac{1}{\widehat{N}_d^2} \sum_{j=1}^{n_d} w_{dj}(w_{dj}-1)\left(y_{dj} - \widehat{\overline{Y}}_d^{dir}\right)^2.$$

This formula is taken from Särndal et al. (1992) (cf. pages 43, 185, 391) and it is valid under the simplifications $w_{dj} = 1/\pi_{dj}$, $\pi_{dj,dj} = \pi_{dj}$ and $\pi_{di,dj} = \pi_{di}\pi_{dj}$ for $i \neq j$, where $\pi_{dj}$ denotes the inclusion probability that the unit $j$ in area $d$ will be in the sample and $\pi_{di,dj}$ denotes the second-order inclusion probability that the both units $i$ and $j$ from area $d$ will be in the sample.

Now we could again describe models incorporating (in addition to sampling weights) some auxiliary data, define composite estimators or deal with design and properties of more complex sampling schemes. But since the design-based approach is not the main concern of this thesis and all the notions necessary for the next chapters were already discussed, we end the brief overview of design-based methods here. For a more detailed discussion of design-based methods, we refer the reader to the article by Lehtonen and Veijanen (2009) which contains a comprehensive review of these methods in small area estimation. Another historical survey of design-based estimators with many references is given in Marker (1999).

Let us finish this section by noting an important disadvantage of the design-based small area estimation. Namely, these methods cannot be used for estimation of small area parameters for areas with no sample. However, in practice it is often the case that only some areas are sampled and estimation is required for all of them, whether sampled or not. This disadvantage can be to some extent overcome by the model-based approach presented in the next section and following chapters.

## 1.2 Model-based small area estimation

The model-based approach treats the population values $y_{dj}$, $d = 1, \ldots, D$, $j = 1, \ldots N_d$, as realizations of random variables $Y_{dj}$ unlike the design-based approach where they were fixed. Relationships among the random variables are expressed by a model of their joint probability distribution and inferences are made with respect to this model. So the bias and variance of some estimate $\widehat{T}$ of a quantity $T$ is under this approach given by

$$\mathrm{E}\left(\widehat{T} - T\right) \qquad \text{and} \qquad \mathrm{V}\left(\widehat{T}\right) = \mathrm{E}\left(\widehat{T} - \mathrm{E}(\widehat{T})\right)^2,$$

respectively, where E denotes the expectation with respect to the underlying model. An estimator $\widehat{T}$ is said to be model-unbiased if $\mathrm{E}(\widehat{T} - T) = 0$. In the following chapters we will often deal with another measure of inaccuracy of the estimator $\widehat{T}$. It is called *mean squared error* (MSE) and its definition is

$$\mathrm{MSE}(\widehat{T}) = \mathrm{E}\left(\widehat{T} - T\right)^2.$$

One can observe that $\mathrm{MSE}(\widehat{T})$ reduces to the variance of the estimation error $\mathrm{V}(\widehat{T} - T)$ if $\widehat{T}$ is model-unbiased estimator of $T$.

After selecting and observing a sample $s$, we will know the realizations $y$'s for the sample units, but the $Y$ values for non-sample units remains unknown. Estimating function

$$h(y_{d1}, \ldots, y_{dN_d})$$

thus entails predicting a function of the unobserved random variables $Y$'s. Notice that the term "prediction" is now used instead of estimation because the target characteristics are generally random under the model. Prediction is thus not used in the usual sense of forecasting future values, but in the sense of making a statistical guess about the unobserved random variables $Y$'s.

**Remark 1** Let us note that in order to follow the notation of the presented papers our notation used in the sequel will not strictly distinguish the random variable $Y$ from its realization $y$. It means that the symbol $y$ may stay for a random variable as well as for a realization of a random variable. In any case the actual meaning will be clear from the context.

The application of explicit models has become very popular and useful in small area estimation since it gives an idea how the data are generated and how different sources of information are combined. This approach has several advantages: 1) it allows formal model building process based on the sample data; 2) "optimal" predictors can be derived under the assumed model; 3) it provides the possibility of expressing uncertainty of the constructed predictors under the assumption that the working model is reasonable; 4) a variety of models can be applied depending on the nature of response variables and complexity of data structures (e.g. time dependance, spatial dependance etc.). Since our conclusions will be based on our model a careful model selection and model diagnostic is an important part of the estimation process. Let us note, however, that the diagnostic of a model may be a difficult problem since SAE models contains assumptions on unobservable random effects which are therefore difficult to verify.

Of a particular attention in SAE are the mixed effects models (models which contain fixed and also random effects) which are very flexible in combining different sources of information. Mixed models typically include area-specific random effect that helps to explain the between area variability in the data which is not explained by the fixed effect part of the model. This is in contrast with the synthetic estimation presented in the previous chapter where the used

implicit regression model assumes no between area variations other than those explained by the auxiliary variables.

Estimation of mean squared errors is an essential part of the small area estimation theory. In the case of estimating population parameters with model-based procedures, this problem has been studied and solved by using empirical best linear unbiased predictors (EBLUPs) of linear parameters under linear models with block-diagonal covariance matrix. Even in this case, the standard estimator is not perfect because it estimates an approximation of the MSE and has a cumbersome expression that needs to be derived for each considered model. For linear mixed models with complicated covariance structure or for generalized linear mixed models, resampling methods and specially bootstrap represent a good alternative because they are efficient and easy to implement. More details about MSE estimation are given in the next chapters.

SAE models are generally classified into two classes based on the data availability of the response and auxiliary variables of interest. If the response variable is available only at small area level we speak about *area level models*. In this case, area level auxiliary information

$$\boldsymbol{x}_d = (x_{d1}, \ldots, x_{dp})^T, \qquad d = 1, \ldots, D,$$

is used. If the response variable is available at the unit level we speak about *unit level models*. In this case unit level auxiliary information

$$\boldsymbol{x}_{dj} = (x_{dj1}, \ldots, x_{djp})^T, \qquad d = 1, \ldots, D,\ j = 1, \ldots, N_d,$$

as well as area level auxiliary information may be used.

In the next chapters we present three models which are in common use in small area estimation. Namely, Fay-Herriot model, nested error regression model and logistic regression model. In fact we can say that most of the recent developments in SAE were connected to these models or to their extensions. We also try to explain what is the contribution of the author to SAE methodology for each of the mentioned model.

# Chapter 2

# Linear mixed models

Since the theory of linear mixed models plays a crucial role in SAE model-based approach, before starting with SAE models in this chapter we first present a definition of linear mixed model and a brief overview of corresponding methods which will be needed for explanation of SAE models for continuous data. Specifically, we present two methods for estimation of parameters of linear mixed models and two approaches for prediction of linear combination of model's effects or linear combinations of observations $\boldsymbol{y}$. Finally we describe the problem of estimating mean square error of linear predictors under linear mixed models.

Let us start with definition of a classical linear regression model which can be expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}\,,$$

where $\boldsymbol{y}$ is a vector of observations, $\boldsymbol{X}$ is a matrix of known covariates, $\boldsymbol{\beta}$ is a vector of unknown regression coefficients and $\boldsymbol{e}$ is a vector of random errors. This model considers the regression coefficients $\boldsymbol{\beta}$ as fixed values. However in some cases it makes sense to assume that some of these coefficients are random. This happens typically when the observations are correlated which may be the case of small area estimation problems where the observations in different small areas are usually assumed to be independent but observation within an area are assumed to be correlated.

A general linear mixed model may be expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}\,, \tag{2.1}$$

where $\boldsymbol{y}_{n\times 1}$ is again a vector of observations, $\boldsymbol{\beta}_{p\times 1}$ is a vector of regression coefficients, which are often called fixed effects, $\boldsymbol{u}_{q\times 1}$ is a vector of random effects, $\boldsymbol{X}_{n\times p}$ and $\boldsymbol{Z}_{n\times q}$ are known matrices of full rank and $\boldsymbol{e}_{n\times 1}$ is a vector of random errors. It is usually assumed that the random effects and random errors are independent, normally distributed with zero means and known variance-covariance matrices,

$$\mathrm{V}(\boldsymbol{u}) = \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^T) = \boldsymbol{V}_u \quad \text{and} \quad \mathrm{V}(\boldsymbol{e}) = \mathrm{E}(\boldsymbol{e}\boldsymbol{e}^T) = \boldsymbol{V}_e$$

which depend on some parameters $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)^T$ called variance components. From (2.1) it follows that the variance-covariance matrix of the vector $\boldsymbol{y}$ has the form

$$\boldsymbol{V} = \mathrm{V}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{V}_u\boldsymbol{Z}^T + \boldsymbol{V}_e. \tag{2.2}$$

It is assumed that the matrices $\boldsymbol{V}, \boldsymbol{V}_u$ and $\boldsymbol{V}_e$ are nonsingular for all possible values of $\boldsymbol{\sigma}$.

## 2.1 Estimation of parameters of linear mixed models

In the model (2.1) there are two unknown vectors of parameters, namely the regression coefficients $\boldsymbol{\beta}$ and the variance components $\boldsymbol{\sigma}$. Let us denote $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\sigma}^T)$ the whole vector of parameters with $p + m$ elements. We describe two methods for estimating $\boldsymbol{\theta}$.

### 2.1.1 MLE

Under normality, which is assumed in our model, the maximum likelihood estimates (MLE) are efficient estimates of the parameter $\boldsymbol{\theta}$. The MLE $\widehat{\boldsymbol{\theta}}$ is defined as the argument of the maxima of the log-likelihood function which for the model (2.1) is given by

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\beta}, \boldsymbol{\sigma}) = c - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\,, \tag{2.3}$$

where $c$ denotes a generic constant. This function must be maximized numerically, e.g. by the well known Fisher-scoring algorithm. From (2.3) it is not difficult to obtain the vector of scores

$$\boldsymbol{s}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left( \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_{p+m}} \right)^T$$

and the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta})$ with the elements

$$\boldsymbol{I}_{i,j}(\boldsymbol{\theta}) = -\mathrm{E}\left( \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right), \qquad i, j = 1, \ldots, p+m\,.$$

The MLE $\widehat{\boldsymbol{\theta}}$ can be then obtained iteratively using the Fisher-scoring updating equation

$$\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i + \boldsymbol{I}(\boldsymbol{\theta}^i)^{-1} \boldsymbol{s}(\boldsymbol{\theta}^i)\,, \tag{2.4}$$

where $\boldsymbol{\theta}^i$ denotes the value of parameter $\boldsymbol{\theta}$ obtained at the $i$th iteration of the algorithm.

### 2.1.2 REML

It is known that the MLE of the variance components $\boldsymbol{\sigma}$ are generally biased. A method called restricted maximum likelihood (REML) was proposed to reduce the bias of the MLE estimators. The idea is to transform the vector $y$ so that the distribution of the transformed vector does not depend on $\beta$ and to estimate the $\boldsymbol{\sigma}$ and $\boldsymbol{\beta}$ independently. For this purpose, the REML method uses the transformed data $\boldsymbol{y}^* = \boldsymbol{A}^T\boldsymbol{y}$ where $\boldsymbol{A}$ is any $n \times (n-p)$ matrix of full rank such that $\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{0}$. The log-likelihood function of the vector $\boldsymbol{y}^*$ is called restricted log-likelihood function and is given by (cf. p. 103 in Rao and Molina (2015))

$$\ell_R(\boldsymbol{\sigma}) = c - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}\,, \tag{2.5}$$

where

$$\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\,.$$

The REML estimate $\widehat{\boldsymbol{\sigma}}$ of $\boldsymbol{\sigma}$ can be again obtained by the updating formula (2.4) but using the vector of scores and Fisher information matrix calculated for the restricted log-likelihood function $\ell_R$. The REML estimate of $\boldsymbol{\beta}$ can be then obtained as

$$\boldsymbol{\beta} = \left(\boldsymbol{X}^T\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{V}}^{-1}\boldsymbol{y}\,,$$

where $\widehat{\boldsymbol{V}}$ denotes the matrix $\boldsymbol{V}$ evaluated at $\widehat{\boldsymbol{\sigma}}$. Note that the REML estimates do not depend on the choice of matrix $\boldsymbol{A}$. For more details and asymptotic distribution of REML estimates see e.g. Cressie and Lahiri (1993).

## 2.2 Prediction of linear combination of model effects

In this section we are interested in estimating a linear combination, $\tau = \boldsymbol{l}^T\boldsymbol{\beta} + \boldsymbol{m}^T\boldsymbol{u}$, of the fixed effects $\beta$ and the realization of random effects $\boldsymbol{u}$, for specified vectors of constants $\boldsymbol{l}$ and $\boldsymbol{m}$. A well known method for the prediction of $\tau$ is the *best linear unbiased prediction* (BLUP) which was given originally by Henderson (1950). The BLUP $\widehat{\tau}^{blup}$ of $\tau$ is in fact linear (in observations $y$) predictor of the form $\widehat{\tau} = \boldsymbol{a}^T\boldsymbol{y} + b$ where constants $\boldsymbol{a}$ and $b$ are determined so that $\widehat{\tau}^{blup}$ is model-unbiased and it minimizes the mean squared error in the class of linear unbiased predictors $\widehat{\tau}$.

If the variance components $\boldsymbol{\sigma}$ of the model (2.1) are known, it can be shown that the BLUP of $\tau$ is given by

$$\widehat{\tau}^{blup} = \boldsymbol{l}^T\widehat{\boldsymbol{\beta}} + \boldsymbol{m}^T\boldsymbol{V}_u\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}),$$

where

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$$

is the least squares estimator of $\boldsymbol{\beta}$ which is sometimes called best linear unbiased estimator (BLUE) under the present setup. By an appropriate choice of the vectors $\boldsymbol{l}$ and $\boldsymbol{m}$ we can immediately obtain the BLUP of $\boldsymbol{u}$ in the form

$$\widehat{\boldsymbol{u}} = \boldsymbol{V}_u\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\,. \tag{2.6}$$

For more details see the chapter 7 in Searle et al. (1992) where it is also shown that the predictor (2.6) is "optimal" in the sense that it minimizes $\mathrm{E}[(\widehat{\boldsymbol{u}} - \boldsymbol{u})^T\boldsymbol{A}(\widehat{\boldsymbol{u}} - \boldsymbol{u})]$ for arbitrary matrix $\boldsymbol{A}$ which is positive definite.

The predictor $\widehat{\tau}^{blup}$ depends on the vector of variance components $\boldsymbol{\sigma}$ which is is typically unknown in practice. Replacing $\boldsymbol{\sigma}$ by an estimator $\widehat{\boldsymbol{\sigma}}$ in the formula of $\widehat{\tau}^{blup}$ we obtain new predictor which is often called *empirical best linear unbiased predictor* (EBLUP) and denoted by $\widehat{\tau}^{eblup}$. Kackar and Harville (1981) give conditions on the estimator $\widehat{\boldsymbol{\sigma}}$ under which the resulting EBLUP remains unbiased. Let us note that the MLE and REML estimators satisfy these conditions under the model (2.1).

## 2.3 Prediction of linear combination of observations $y$

Let us now consider a finite population of $N$ elements with population vector $\boldsymbol{y} = (y_1, \ldots, y_N)$ following the model introduced in (2.1) with population sizes $N$ in the place of sizes $n$. From the population a sample of size $n$ is selected. Without loss of generality we can reorder the population so that $\boldsymbol{y} = (\boldsymbol{y}_s^T, \boldsymbol{y}_r^T)^T$, where $\boldsymbol{y}_s$ is the vector of $n$ observed elements and $\boldsymbol{y}_r$ is the vector of $N - n$ unobserved elements. In the following, the index $s$ for the sample and the index $r$ for the rest of the population will be used when appropriate. In this notation and taking into account the reordering we can write the matrix $\boldsymbol{X}$ and the covariance matrix $\boldsymbol{V}$ in the block form

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_s \\ \boldsymbol{X}_r \end{pmatrix}, \qquad \boldsymbol{V} = \mathrm{V}(\boldsymbol{y}) = \begin{pmatrix} \boldsymbol{V}_{ss} & \boldsymbol{V}_{sr} \\ \boldsymbol{V}_{rs} & \boldsymbol{V}_{rr} \end{pmatrix}, \tag{2.7}$$

where $\boldsymbol{V}_{ss} = \mathrm{V}(\boldsymbol{y}_s)$ is the covariance matrix of the observed elements, $\boldsymbol{V}_{rr} = \mathrm{V}(\boldsymbol{y}_r)$ is the covariance matrix of the unobserved elements and $\boldsymbol{V}_{rs} = \mathrm{Cov}(\boldsymbol{y}_r, \boldsymbol{y}_s)$. We assume that $\boldsymbol{V}_{ss}$ is positive definite.

On the basis of the sample $\boldsymbol{y}_s$, we are interested in the estimation of a linear combination

$$\eta = \boldsymbol{a}^T\boldsymbol{y}$$

for a given vector $\boldsymbol{a} = (a_1, \ldots, a_N)$. If, for example, each $a_i = 1/N$, then the target variable is the population mean. The vector $\boldsymbol{a}$ can also be partitioned into parts, $\boldsymbol{a} = (\boldsymbol{a}_s^T, \boldsymbol{a}_r^T)^T$, corresponding to the sample and non-sample units and our estimation target can be expressed in the form

$$\eta = \boldsymbol{a}^T \boldsymbol{y} = \boldsymbol{a}_s^T \boldsymbol{y}_s + \boldsymbol{a}_r^T \boldsymbol{y}_r \,.$$

From the last formula it is clear that the problem of estimating $\boldsymbol{a}^T \boldsymbol{y}$ is equivalent to that of predicting the linear combination $\boldsymbol{a}_r^T \boldsymbol{y}_r$ of the unobserved random variables.

As $\eta$ is a linear parameter, the predictor minimizing the mean squared error in the class of model-unbiased predictors is the BLUP. From the general prediction theorem (see e.g. Section 2.2 of Valliant et al. (2000)), it follows that the BLUP of $\eta$, under the model (2.1), is

$$\widehat{\eta}^{blup} = \boldsymbol{a}_s^T \boldsymbol{y}_s + \boldsymbol{a}_r^T \left[ \boldsymbol{X}_r \widehat{\boldsymbol{\beta}} + \boldsymbol{V}_{rs} \boldsymbol{V}_{ss}^{-1} (\boldsymbol{y}_s - \boldsymbol{X}_s \widehat{\boldsymbol{\beta}}) \right] , \tag{2.8}$$

where

$$\widehat{\boldsymbol{\beta}} = \left( \boldsymbol{X}_s^T \boldsymbol{V}_{ss} \boldsymbol{X}_s \right)^{-1} \boldsymbol{X}_s^T \boldsymbol{V}_{ss}^{-1} \boldsymbol{y}_s \,. \tag{2.9}$$

The EBLUP $\widehat{\eta}^{eblup}$ is again obtained by substituting an estimator $\widehat{\boldsymbol{\sigma}}$ in the formula (2.8).

## 2.4 Mean squared error of EBLUP

Estimation of the mean squared error of EBLUP is of significant interest, since in practical applications we need an estimator of MSE as a measure of variability associated with our predictor. While EBLUP is quite easy to obtain, estimation of its MSE is a challenging problem. Let us illustrate the basic principles for the EBLUP $\widehat{\eta}^{eblup}$ predicting the linear combination of observations $y$. For the EBLUP $\widehat{\tau}^{eblup}$ used for prediction of the linear combination of the model effects the arguments are similar.

The mean squared error of the EBLUP $\widehat{\eta}^{eblup}$, denoted by

$$MSE(\widehat{\eta}^{eblup}) = \mathrm{E} \left( \widehat{\eta}^{eblup} - \eta \right)^2 ,$$

can be decomposed as

$$MSE(\widehat{\eta}^{eblup}) = MSE(\widehat{\eta}^{blup}) + \mathrm{E} \left( \widehat{\eta}^{eblup} - \widehat{\eta}^{blup} \right)^2 , \tag{2.10}$$

where the first term is the MSE of the BLUP $\widehat{\eta}^{blup}$. It is clear that the MSE of the EBLUP is always larger than the MSE of BLUP and the increase is caused by the variability of the estimator $\widehat{\boldsymbol{\sigma}}$ used for calculating the EBLUP. While the first term of the decomposition (2.10) can be evaluated, the second term is generally intractable and must be approximated.

Kackar and Harville (1984) provided first simplification of the MSE and proposed an estimator based on it. But the accuracy of the approximation was not studied. In a pioneering work, Prasad and Rao (1990) gave a new approximation for models with block-diagonal covariance matrices. They also studied a new estimator of the MSE and gave the specific expressions of this estimator for some concrete models. The conditions imposed on the estimators of the variance components are satisfied by estimators obtained by the Fitting Constants Method, also called Henderson method 3, but they cannot be verified for maximum likelihood estimators. Datta and Lahiri (2000) provided MSE estimators for general models with block-diagonal covariance matrices, when variance components are estimated by MLE or REML methods. Das et al. (2004) studied the approximation of the MSE for a wider class of models when variance components are estimated by MLE or REML.

Here we for illustration present a general formula for approximation of the MSE of $\widehat{\eta}^{eblup}$ obtained by following Prasad and Rao (1990) and Das et al. (2004). Using the notation of Section 2.3, the approximation can be expressed in the following way:

$$MSE(\widehat{\eta}^{eblup}) = g_1(\boldsymbol{\sigma}) + g_2(\boldsymbol{\sigma}) + g_3(\boldsymbol{\sigma}) + g_4(\boldsymbol{\sigma}), \tag{2.11}$$

where

$$\begin{aligned}
g_1(\boldsymbol{\sigma}) &= \boldsymbol{a}_r^T \boldsymbol{Z}_r \boldsymbol{T}_s \boldsymbol{Z}_r^T \boldsymbol{a}_r, \\
g_2(\boldsymbol{\sigma}) &= [\boldsymbol{a}_r^T \boldsymbol{X}_r - \boldsymbol{a}_r^T \boldsymbol{Z}_r \boldsymbol{T}_s \boldsymbol{Z}_s^T \boldsymbol{V}_{e,ss}^{-1} \boldsymbol{X}_s] \boldsymbol{Q}_s [\boldsymbol{X}_r^T \boldsymbol{a}_r - \boldsymbol{X}_s^T \boldsymbol{V}_{e,ss}^{-1} \boldsymbol{Z}_s \boldsymbol{T}_s \boldsymbol{Z}_r^T \boldsymbol{a}_r], \\
g_3(\boldsymbol{\sigma}) &\approx \operatorname{tr}\left\{ (\nabla \boldsymbol{b}^T) \boldsymbol{V}_{ss} (\nabla \boldsymbol{b}^T)^T E\left[ (\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma})(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma})^T \right] \right\}, \\
g_4(\boldsymbol{\sigma}) &= \boldsymbol{a}_r^T \boldsymbol{V}_{e,rr} \boldsymbol{a}_r,
\end{aligned}$$

and $\boldsymbol{T}_s = \boldsymbol{V}_u - \boldsymbol{V}_u \boldsymbol{Z}_s^T \boldsymbol{V}_{ss}^{-1} \boldsymbol{Z}_s \boldsymbol{V}_u$, $\boldsymbol{Q}_s = (\boldsymbol{X}_s^T \boldsymbol{V}^{-1} \boldsymbol{X}_s)^{-1}$, $\boldsymbol{b}^T = \boldsymbol{a}_r^T \boldsymbol{Z}_r \boldsymbol{V}_u \boldsymbol{Z}_s^T \boldsymbol{V}_{ss}^{-1}$. The symbols $\boldsymbol{V}_{e,ss}$, $\boldsymbol{Z}_s$ and others are used to denote the observed elements of the matrices $\boldsymbol{V}_e$, $\boldsymbol{Z}$, correspondingly to the notation introduced in (2.7).

The Prasad-Rao estimator of $MSE(\widehat{\eta}^{eblup})$ is then defined as

$$mse(\widehat{\eta}^{eblup}) = g_1(\widehat{\boldsymbol{\sigma}}) + g_2(\widehat{\boldsymbol{\sigma}}) + 2g_3(\widehat{\boldsymbol{\sigma}}) + g_4(\widehat{\boldsymbol{\sigma}}),$$

where $\widehat{\boldsymbol{\sigma}}$ is REML estimator of $\boldsymbol{\sigma}$. Notice that a coefficient 2 has appeared before the term $g_3(\hat{\boldsymbol{\sigma}})$. This coefficient represents correction of bias introduced by substituting $\boldsymbol{\sigma}$ by its estimate $\widehat{\boldsymbol{\sigma}}$ in the formula (2.11).

# Chapter 3

# Models for continuous responses

We are now ready to return to the problem of small area estimation and in this chapter we deal with models for continuous responses. As mentioned before, these models may be divided into area level and unit level models which are treated in separate sections.

## 3.1 Area level models

The basic area level model for continuous responses was firstly formulated by Fay and Herriot (1979) and was used to improve the information obtained from design-based small area estimates by using some additional auxiliary data $\boldsymbol{x}_d = (x_{d1}, \ldots, x_{dp})$ at area level. It can be described in the following way.

Let $\mu_d$ denote the characteristic of interest in the area $d$, e.g. the area population mean

$$\mu_d = \overline{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}, \qquad d = 1, \ldots, D\,,$$

and $y_d$ be a direct estimator of $\mu_d$ with known design-based variance $\sigma_d^2$. The Fay-Herriot model is composed of two levels:

- Sampling model:
$$y_d = \mu_d + e_d, \qquad d = 1, \ldots, D\,,$$
- Linking model:
$$\mu_d = \boldsymbol{x}_d^T \boldsymbol{\beta} + u_d, \qquad d = 1, \ldots, D\,,$$

where $e_d \sim N(0, \sigma_d^2)$ are independent sampling errors, $\boldsymbol{\beta}$ is a vector of regression coefficients and $u_d \sim N(0, \sigma_u^2)$ are random effects (model errors) which are assumed to be independent and identically distributed (i.i.d.) and independent of the sampling errors $e_d$.

In the above described model, the sampling model is used to account for the sampling variability of the direct estimates $y_d$. The linking model links the true small area parameters $\mu_d$ to a vector of $p$ known auxiliary variables. The parameters $\boldsymbol{\beta}$ and $\sigma_u^2$ of the linking model are generally unknown and are estimated from the available data.

Let us note that the Fay-Herriot model can be expressed as an area level linear mixed model

$$y_d = \boldsymbol{x}_d^T \boldsymbol{\beta} + u_d + e_d, \qquad d = 1, \ldots, D\,, \tag{3.1}$$

where $u_d \sim N(0, \sigma_u^2)$, $e_d \sim N(0, \sigma_d^2)$, $d = 1, \ldots, D$; they are all mutually independent and the variances $\sigma_1^2, \ldots, \sigma_D^2$ are known. Our task is now to estimate the quantity $\mu_d = \boldsymbol{x}_d^T \boldsymbol{\beta} + u_d$ with the aim to improve the direct estimates $y_d$. In fact, we would like to eliminate the sampling error contained in the direct estimates.

**Remark 2** Fay and Herriot (1979) originally used their model for estimating per-capita income (PCI) of small places with population less than 1000 in the United States. The objective was to decrease the variability of the design-based PCI estimates obtained by the U.S. Census Bureau in 1970. As auxiliary data they used tax-refund data for 1969 and data on housing from the 1970 census. The random effects $u_d$ were used to capture the additional area-specific effects which were not explained by the area level auxiliary variables. Fay-Herriot model demonstrated that it can provide EBLUP estimators with better performance than the direct survey estimator and a synthetic estimator used before by the U.S. Census Bureau.

### 3.1.1 Empirical best predictors

To derive the EBLUP of the parameter of interest notice that the Fay-Herriot model (3.1) can be written in the matrix form

$$\begin{pmatrix} y_1 \\ \vdots \\ y_D \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{D1} & \cdots & x_{dp} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_D \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_D \end{pmatrix}$$

corresponding to the general formula of linear mixed model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$ given in (2.1). So the theory reviewed in the previous chapter can be used for estimating parameters of the model and for constructing EBLUP and its MSE estimates.

Taking into account that matrix $\boldsymbol{Z}$ is the identity matrix $\boldsymbol{I}_{D\times D}$ in the present model and the matrix $V$ can be expressed as

$$\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{V}_u\boldsymbol{Z}^T + \boldsymbol{V}_e = \boldsymbol{V}_u + \boldsymbol{V}_e = \begin{pmatrix} \sigma_u^2 + \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_u^2 + \sigma_D^2 \end{pmatrix} = \operatorname*{diag}_{1\le d\le D} (\sigma_u^2 + \sigma_d^2)\,,$$

we can use the formula (2.6) to derive the BLUP of the components of vector $\boldsymbol{u}$ in the form

$$\widehat{u}_d = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_d^2} \left(y_d - \boldsymbol{x}_d^T\widehat{\boldsymbol{\beta}}\right), \qquad d = 1, \ldots, D\,,$$

where $\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$ is the BLUE estimator of $\beta$. The BLUP of the parameter $\mu_d$ is thus

$$\widehat{\overline{Y}}_d^{blup} = \widehat{\mu}_d = \boldsymbol{x}_d^T\widehat{\boldsymbol{\beta}} + \widehat{u}_d = \boldsymbol{x}_d^T\widehat{\boldsymbol{\beta}} + \frac{\sigma_u^2}{\sigma_u^2 + \sigma_d^2}\left(y_d - \boldsymbol{x}_d^T\widehat{\boldsymbol{\beta}}\right) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_d^2}\, y_d + \frac{\sigma_d^2}{\sigma_u^2 + \sigma_d^2}\, \boldsymbol{x}_d^T\widehat{\boldsymbol{\beta}}\,. \quad (3.2)$$

The EBLUP $\widehat{\overline{Y}}_d^{eblup}$ is then obtained by substituting an estimator $\widehat{\sigma}_u^2$ of the parameter $\sigma_u^2$ into the formula (3.2).

### 3.1.2 MSE of EBLUP

Following the steps described in the Section 2.4, Prasad and Rao (1990) derived an approximation of the mean squared error of the EBLUP of $\overline{Y}_d$ under Fay-Herriot model. The approximation is

$$MSE(\widehat{\overline{Y}}_d^{eblup}) \approx g_{1d}(\sigma_u^2) + g_{2d}(\sigma_u^2) + g_{d3}(\sigma_u^2),$$

where

$$\begin{aligned} g_{1d}(\sigma_u^2) &= \frac{\sigma_u^2\sigma_d^2}{\sigma_u^2+\sigma_d^2}, \\ g_{2d}(\sigma_u^2) &= \frac{\sigma_d^4}{(\sigma_u^2+\sigma_d^2)^2}\,\boldsymbol{x}_d^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_d, \\ g_{3d}(\sigma_u^2) &= \frac{\sigma_d^4}{(\sigma_u^2+\sigma_d^2)^3}\,var(\widehat{\sigma}_u^2)\,, \end{aligned}$$

and $var(\widehat{\sigma}_u^2)$ is the asymptotic variance of the estimator $\widehat{\sigma}_u^2$ which must be evaluated on the basis of the method used for estimation of $\sigma_u^2$. If e.g. maximum likelihood method is used for estimating the model parameters then the asymptotic variance of $\sigma_u^2$ can be approximated by the corresponding diagonal element of the inverse Fisher information matrix. Cressie and Lahiri (1993) give the asymptotic variance formula also for REML estimators.

If we now substitute an estimate $\widehat{\sigma}_u^2$ instead of $\sigma_u^2$ in the formula of MSE, the resulting estimator will be biased. The following approximately unbiased formula can be used for calculating estimator $mse$ of the mean squared error $MSE$,

$$mse(\widehat{\overline{Y}}_d^{eblup}) = g_{1d}(\widehat{\sigma}_u^2) + g_{2d}(\widehat{\sigma}_u^2) + 2g_{3d}(\widehat{\sigma}_u^2).$$

For more details see Prasad and Rao (1990).

### 3.1.3 Contribution of the author to area level models

Many different extensions of Fay-Herriot model have been proposed in the literature. For example, a multivariate generalization was studied by González-Manteiga et al. (2008b). Models assuming spatial correlation between neighboring areas were considered in Singh et al. (2005), Petrucci and Salvati (2006), and Pratesi and Salvati (2008) between others. Models with temporal correlation, using data from different time instants to improve the estimator at the current instant, have been proposed e.g. in Choudry and Rao (1989), Rao and Yu (1994) or Ghosh et al. (1996).

In this section we will try to explain the contribution of the author to the field of area level models.

**I)** The Fay-Herriot model typically assumes that the domain random effects have a common constant variance. However, in practise we often encounter situations where the domains are divided in two groups and the direct estimates have different behaviour within them. This situation may happen if we are interested in producing estimates for domains constructed by crossing geographical area with sex category. In the paper

> Esteban, M.D., Herrador, M., Hobza, T., Morales, D. (2011). A Fay-Herriot model with different random effect variances. *Communications in Statistics – Theory and Methods*, 40(5), pp. 785-797,

presented in Section 5.1 on page 51, we suppose that the domains are divided into two groups, denoted $A$ and $B$, and we propose the following modification of the Fay-Herriot model

$$y_d = \boldsymbol{x}_d^T\boldsymbol{\beta} + u_d + e_d, \quad d = 1,\ldots,D = D_A + D_B, \tag{3.3}$$

where

$$u_1,\ldots,u_{D_A} \sim N(0,\sigma_A^2), \qquad u_{D_A+1},\ldots,u_D \sim N(0,\sigma_B^2),$$

$e_d \sim N(0, \sigma_d^2)$, $d = 1, \ldots, D$; they are all mutually independent and the variances $\sigma_1^2, \ldots, \sigma_D^2$ are known. That means we assume that the variances $\sigma_A^2$, $\sigma_B^2$ of the random effects may be different in groups $A$ and $B$.

In the paper we give formulas to estimate the model parameters, to calculate EBLUPs and to estimate their means squared errors. Two simulation experiments, at area level and at unit level, are presented and they show that if the proposed model is true and the standard Fay-Herriot model is used, then a lack of precision is achieved. A motivating application to real data from the Spanish Labour Force Survey is also given. The application deals with estimating proportion of unemployed people by sex in the Canary Islands in the second trimester of 2003. As auxiliary variables the population means of age and work categories were used. The conclusion from the results of the application is that the newly introduced EBLUP gives better results than those obtained by applying the standard Fay-Herriot model to the whole set of direct estimates or those obtained by applying by sex two independent standard Fay-Herriot models. For more details see Section 5.1.

**II)** Our second modification of Fay-Herriot model consists in allowing the area effects $u_d$ to be fixed for some of the domains. The considered model thus has both fixed and random area effects. The model is suitable for data sets containing some domains where the direct estimates are e.g. much larger than in the rest or where in some domains the direct estimates have been obtained with large sample sizes and therefore they are reliable. In the latter case, an appealing property of the modified model is that the direct estimates coincide with EBLUP estimates in the selected domains with fixed area effects. The model was formulated in the paper

> Herrador, M., Esteban, M. D., Hobza, T., Morales, D. (2011). An Area-Level Model with Fixed or Random Domain Effects in Small Area Estimation Problems. *Modern Mathematical Tools and Techniques in Capturing Complexity - Understanding Complex Systems*, Springer Berlin, pp. 303 - 314,

presented in Section 5.2 on page 65. It can be written in terms of fixed effect ($F$) part and random effect ($R$) part in the following way

$$(F) \qquad y_d = x_d^T \boldsymbol{\beta} + \mu_d + e_d, \quad d = 1, \ldots, D_F,$$

$$(R) \qquad y_d = x_d^T \boldsymbol{\beta} + u_d + e_d, \quad d = D_F + 1, \ldots, D,$$

where $\mu_1, \ldots, \mu_{D_F}$ are the unknown parameters corresponding to the fixed effect levels and $u_{D_F+1}, \ldots, u_D$ are i.i.d. $N(0, \sigma_u^2)$ distributed random variables independent of the random errors $e_d$.

In the paper algorithms to fit the model, to calculate EBLUP and to estimate its MSE are derived. The properties of the proposed estimator are studied by a simulation experiment showing that the EBLUP and its MSE estimates are more precise than the estimates obtained under the classical Fay-Herriot model if there are some domains with different behavior of the direct estimates.

The model was also applied to the real data problem described in the previous paragraph. In the Canary Islands 2013 data there are two domains with much larger sample sizes than the rest so these two domains were put to the fixed part of the model. The application showed that although the sample sizes are very different, the proportions of unemployed people behave similarly across all domains so almost no differences between the proposed and classical model were observed. Nevertheless, the possibility of obtaining model-based estimates that coincide with the direct ones in the fixed part of the model is attractive from the point of view of modelers and official statisticians.

## 3.2 Unit level models

If the response variables and auxiliary data are available not only at the area level but also at the individual level, unit level linear mixed models may be used to estimate domain parameters. These models typically assume that the regression coefficients are constant but the intercepts are random with realizations on domains. Such random intercept models in fact assign a regression line with the same slope but different intercepts to each domain. The random intercept variance then refers to the variability of the line heights at the origin.

In the setup of small area estimation the first model of this type was proposed by Battese et al. (1988). Their model is usually called *"nested error regression model"* and its definition is

$$y_{dj} = \boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d + e_{dj}, \qquad d = 1,\dots,D, \quad j = 1,\dots,n_d\,, \tag{3.4}$$

where $y_{dj}$ is the $j$th observation from area $d$, $\boldsymbol{x}_{dj}$ is a corresponding vector of auxiliary variables, $\boldsymbol{\beta}$ is a vector of unknown regression parameters, $u_d$ is an area-specific random effect (intercept) and $e_{dj}$ is a random error. It is further assumed that the random effects $u_d$'s are *i.i.d.* random variables with $N(0,\sigma_u^2)$ distribution which are independent of the random errors $e_{dj}$'s. The random errors $e_{dj}$'s are also assumed to be independent with the distribution $e_{dj} \sim N(0, w_{dj}^{-1}\,\sigma_e^2)$, where $w_{dj}$'s are known heteroscedasticity weights.

**Remark 3** Battese at al. (1988) used the model (3.4) for estimating the areas under corn and soybeans for each of 12 counties in North Central Iowa. Each county was divided into segments and the sample observations $y_{dj}$, expressing the number of hectares of corn and soybeans, were obtained for a sample of segments by interviewing farmers. Sample sizes for the counties were very small, ranging from 1 to 6 segments, making the direct estimates highly unprecise. In order to increase precision of the estimates and to allow the use of a model, auxiliary variables were obtained from satellite data for all the segments. The meaning of auxiliary variables $\boldsymbol{x}_{dj} = (x_{dj1}, x_{dj2})$ was the number of pixels classified as corn or soybeans in the satellite picture of the concrete segment $j$ in the county $d$.

Notice that the model (3.4) can be written in the matrix form

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \hline y_{21} \\ \vdots \\ y_{2n_2} \\ \hline \vdots \\ \hline y_{D1} \\ \vdots \\ y_{Dn_D} \end{pmatrix} = \begin{pmatrix} x_{111} & \cdots & x_{11p} \\ \vdots & \vdots & \vdots \\ x_{1n_11} & \cdots & x_{1n_1p} \\ \hline x_{211} & \cdots & x_{21p} \\ \vdots & \vdots & \vdots \\ x_{2n_21} & \cdots & x_{2n_2p} \\ \hline \vdots & \vdots & \vdots \\ \hline x_{D11} & \cdots & x_{D1p} \\ \vdots & \vdots & \vdots \\ x_{Dn_D1} & \cdots & x_{Dn_Dp} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ \hline 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ \hline \vdots & \vdots & & \vdots \\ \hline 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_D \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{1n_1} \\ \hline e_{21} \\ \vdots \\ e_{2n_2} \\ \hline \vdots \\ \hline e_{D1} \\ \vdots \\ e_{Dn_D} \end{pmatrix}$$

which again corresponds to the general formula of linear mixed model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$ given in (2.1). Let us note that e.g. the matrix $\boldsymbol{Z}$ can be written as $\boldsymbol{Z} = \mathrm{diag}_{1\le d\le D}(\mathbf{1}_{n_d})$, where $\mathbf{1}_m = (1,\dots,1)^T_{1\times m}$.

### 3.2.1 Empirical best predictors

Let us now consider a finite population of $N_d$ elements in area $d$ from which a sample of sizes $n_d$ is selected. In the following we assume that the model (3.4) holding for the sample data

holds also for the whole population so that there is no selection bias. This implies that the sample design used to select the sample is ignorable and need not to be taken into account in construction of our predictors.

On the basis of the selected sample, we are interested in the estimation of the mean of the small area $d$, i.e.

$$\overline{Y}_d = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj} = \boldsymbol{a}^T\boldsymbol{y},$$

where $\boldsymbol{a}^T = \frac{1}{N_d}\left(\mathbf{0}^T_{N_1},\ldots,\mathbf{0}^T_{N_{d-1}},\mathbf{1}^T_{N_d},\mathbf{0}^T_{N_{d+1}},\ldots,\mathbf{0}^T_{N_D}\right)$ and $\mathbf{0}^T_m = (0,\ldots,0)_{1\times m}$. Taking into account the reordering introduced in Section 2.3 the linear parameter $\boldsymbol{a}^T\boldsymbol{y}$ can be expressed in the form

$$\boldsymbol{a}^T\boldsymbol{y} = \boldsymbol{a}_s^T\boldsymbol{y}_s + \boldsymbol{a}_r^T\boldsymbol{y}_r\,,$$

where

$$\boldsymbol{a}_s^T = \frac{1}{N_d}\left(\mathbf{0}^T_{n_1},\ldots,\mathbf{0}^T_{n_{d-1}},\mathbf{1}^T_{n_d},\mathbf{0}^T_{n_{d+1}},\ldots,\mathbf{0}^T_{n_D}\right)$$

and

$$\boldsymbol{a}_r^T = \frac{1}{N_d}\left(\mathbf{0}^T_{N_1-n_1},\ldots,\mathbf{0}^T_{N_{d-1}-n_{d-1}},\mathbf{1}^T_{N_d-n_d},\mathbf{0}^T_{N_{d+1}-n_{d+1}},\ldots,\mathbf{0}^T_{N_D-n_D}\right).$$

If the variance components $\sigma_u^2,\sigma_e^2$ are known, the general formula (2.8) may be used for calculation the EBLUP of $\boldsymbol{a}^T\boldsymbol{y}$. We are not going to give a complete derivation here, for more details we refer to the papers presented in Sections 5.3 and 5.5, let us just note that e.g. for the matrix $\boldsymbol{V}_{rs}$ it holds (cf. (2.2))

$$\boldsymbol{V}_{rs} = \boldsymbol{Z}_r\boldsymbol{V}_u\boldsymbol{Z}_s^T + \boldsymbol{V}_{e,rs}\,,$$

where $\boldsymbol{Z}_r = \mathrm{diag}_{1\le d\le D}(\mathbf{1}_{N_d-n_d})$ and $\boldsymbol{V}_{e,rs} = \mathbf{0}$ since $e_{dj}$'s are supposed to be independent in the whole population. Substituting these terms together with the known variance-covariance matrices of the vectors $\boldsymbol{u}$ and $\boldsymbol{e}$ into the formula (2.8) and after some straightforward algebra, the expression of the BLUP predictor of $\overline{Y}_d$ takes the form

$$\widehat{\overline{Y}}_d^{\,blup} = (1-f_d)\left[\overline{\boldsymbol{X}}_d\widehat{\boldsymbol{\beta}} + \gamma_d^w\left(\widehat{\overline{Y}}_d^{\,dir} - \widehat{\overline{\boldsymbol{X}}}_d^{\,dir}\widehat{\boldsymbol{\beta}}\right)\right] + f_d\left[\widehat{\overline{y}}_d + (\overline{\boldsymbol{X}}_d - \widehat{\overline{\boldsymbol{X}}}_d)\widehat{\boldsymbol{\beta}}\right], \tag{3.5}$$

where $\widehat{\boldsymbol{\beta}}$ is the BLUE estimator given in (2.9),

$$f_d = \frac{n_d}{N_d},\quad \gamma_d^w = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{w_d}}\quad \text{for}\quad w_d = \sum_{j=1}^{n_d} w_{dj},$$

$$\overline{\boldsymbol{X}}_d = \frac{1}{N_d}\sum_{j=1}^{N_d}\boldsymbol{x}_{dj}^T,\qquad \widehat{\overline{\boldsymbol{X}}}_d = \frac{1}{n_d}\sum_{j=1}^{n_d}\boldsymbol{x}_{dj}^T,\qquad \widehat{\overline{y}}_d = \frac{1}{n_d}\sum_{j=1}^{n_d} y_{dj}$$

and

$$\widehat{\overline{Y}}_d^{\,dir} = \frac{1}{w_d}\sum_{j=1}^{n_d} w_{dj}y_{dj},\qquad \widehat{\overline{\boldsymbol{X}}}_d^{\,dir} = \frac{1}{w_d}\sum_{j=1}^{n_d} w_{dj}\boldsymbol{x}_{dj}^T.$$

The empirical best linear predictor $\widehat{\overline{Y}}_d^{\,eblup}$ of $\overline{Y}_d$ is then obtained from (3.5) by substituting variance components by their consistent estimators.

**Remark 4** Notice, that the matrix $\boldsymbol{X}_r$ which contains the auxiliary variables for the non-sampled elements and appears in the formula (2.8) is not involved in the final formula (3.5) of the EBLUP. Actually, for the assumed form of the vector $\boldsymbol{a}$ the quantity depending on $\boldsymbol{X}_r$ can be expressed as a function of the means $\overline{\boldsymbol{X}}_d$ and $\widehat{\overline{\boldsymbol{X}}}_d$. Thus, it is not necessary to know the auxiliary data $\boldsymbol{x}_{dj}$ for all elements of the population, which would be too restrictive in practical applications. It is enough to know $\boldsymbol{x}_{dj}$ for the units in the sample and in addition to know the population means $\overline{\boldsymbol{X}}_d$. This is the information usually obtained from some external sources.

**Remark 5** Another feature of the EBLUP based on the nested error regression model is that it can provide predictions also for non-sampled areas. Let us assume the there is no sample in area $d$, i.e. no variables $y_{dj}$ were observed for area $d$ so $n_d = 0$ . Then the formula (3.5) for EBLUP reduces to

$$\widehat{\overline{Y}}_d = \overline{\boldsymbol{X}}_d \widehat{\boldsymbol{\beta}} \,.$$

This estimator takes into account the regression parameter $\boldsymbol{\beta}$ estimated from the sampled domains and the auxiliary information for area $d$ in the form of the area population mean $\overline{\boldsymbol{X}}_d$.

### 3.2.2 Mean squared error of EBLUP

Since the assumed model belongs to the class of linear mixed models, we can use the general formula (2.11) and obtain approximation of the mean squared error of the above described EBLUP in the form

$$MSE\left(\widehat{\overline{Y}}_d^{eblup}\right) = g_{1d}(\sigma_e^2, \sigma_u^2) + g_{2d}(\sigma_e^2, \sigma_u^2) + g_{3d}(\sigma_e^2, \sigma_u^2) + g_{4d}(\sigma_e^2, \sigma_u^2)\,, \tag{3.6}$$

where

$$\begin{aligned}
g_{1d}(\sigma_e^2, \sigma_u^2) &= (1-f_d)^2(1-\gamma_d^w)\sigma_u^2 \\
g_{2d}(\sigma_e^2, \sigma_u^2) &= (1-f_d)^2 \left(\overline{\boldsymbol{X}}_d - \gamma_d^w \widehat{\overline{\boldsymbol{X}}}_d^{dir}\right) (\boldsymbol{X}_s^T \boldsymbol{V}_s^{-1} \boldsymbol{X}_s)^{-1} \left(\overline{\boldsymbol{X}}_d - \gamma_d^w \widehat{\overline{\boldsymbol{X}}}_d^{dir}\right)^T \\
g_{3d}(\sigma_e^2, \sigma_u^2) &= (1-f_d)^2 \left(\sigma_u^2 + \frac{\sigma_e^2}{w_d}\right)^{-3} \frac{1}{w_d^2} \left\{\sigma_e^4 \mathrm{V}(\widehat{\sigma}_u^2) - 2\sigma_u^2\sigma_e^2 \mathrm{Cov}(\widehat{\sigma}_u^2, \widehat{\sigma}_e^2) + \sigma_u^4 \mathrm{V}(\widehat{\sigma}_e^2)\right\}, \\
g_{4d}(\sigma_e^2, \sigma_u^2) &= \frac{\sigma_e^2(\mathcal{V}_d - \nu_d)}{N_d^2}\,, \qquad \mathcal{V}_d = \sum_{j=1}^{N_d} w_{dj}^{-1}, \qquad \nu_d = \sum_{j=1}^{n_d} w_{dj}^{-1}\,.
\end{aligned}$$

The asymptotic variances $\mathrm{V}(\widehat{\sigma}_u^2)$, $\mathrm{V}(\widehat{\sigma}_e^2)$ and the asymptotic covariance $\mathrm{Cov}(\widehat{\sigma}_u^2, \widehat{\sigma}_e^2)$ must be evaluated on the basis of the method used for estimation of the variances $\sigma_u^2$, $\sigma_e^2$. Since the derivation of the above presented formulas is quite technical, we present here just these final expressions and for more details we refer the reader to Appendix 2 and Appendix 1 of the paper presented in Section 5.3, where the procedure is illustrated for a more complex model and asymptotic variances for REML estimates of the variance components are given.

Finally, by substituting the REML estimates $\widehat{\sigma}_u^2$, $\widehat{\sigma}_e^2$ into the formula (3.6) and correcting bias, we obtain the estimator $mse$ of the mean squared error

$$mse\left(\widehat{\overline{Y}}_d^{eblup}\right) = g_{1d}(\widehat{\sigma}_e^2, \widehat{\sigma}_u^2) + g_{2d}(\widehat{\sigma}_e^2, \widehat{\sigma}_u^2) + 2\, g_{3d}(\widehat{\sigma}_e^2, \widehat{\sigma}_u^2) + g_{4d}(\widehat{\sigma}_e^2, \widehat{\sigma}_u^2)\,.$$

### 3.2.3 Contribution of the author to unit level models

Let us mention some extensions of the model (3.4) assumed in the literature. Fuller and Harter (1987) propose a multivariate nested error regression model suitable for the cases where the variable of interest $y_{dj}$ is a vector. Stukel (1991) studies two-fold nested error regression models assuming that the small areas are further divided into clusters. Datta and Ghosh (1991) consider a general linear mixed model that includes the model (3.4) as a special case.

Contribution of the author consists in the following three modifications of the nested error regression model.

**I)** In the first contribution,

> Esteban, M.D., Herrador, M., Hobza, T., Morales, D. (2013). A modified nested-error regression model for small area estimation. *Statistics: A Journal of Theoretical and Applied Statistics*, 47(2), pp. 258-273,

presented in Section 5.3 on page 78, we employ the idea already used in the case of Fay-Herriot model and we propose a model having both fixed and random levels which can be written in terms of fixed effect ($F$) part and random effect ($R$) part in the following way:

$$(F) \qquad y_{dj} = x_{dj}^T\boldsymbol{\beta} + \mu_d + e_{dj}, \quad d = 1, \ldots, D_F,\ j = 1, \ldots, n_d,$$

$$(R) \qquad y_{dj} = x_{dj}^T\boldsymbol{\beta} + u_d + e_{dj}, \quad d = D_F + 1, \ldots, D,\ j = 1, \ldots, n_d,$$

where $\mu_1, \ldots, \mu_{D_F}$ are unknown parameters corresponding to the fixed effects and $u_{D_F+1}, \ldots, u_D$ are i.i.d. random effects independent of the random errors $e_{dj}$. This model is useful if there are, for example, some domains where the quantity of interest is higher than in the rest of the domains. We can imagine for instance that the quantity of interest is the personal income and in the considered country there exist some "outlying" domains with much higher income. In such case, the model intercepts may be much higher in the mentioned domains and the traditional model does not fit well to data since some domains are responsible for overestimating the intercept variance which negatively affects the EBLUP estimates. Another interesting case for applying the modified model could be if there is demand, for administrative or political reasons, for increasing the precision of estimates for any given domains. Such domains would be included in the fixed part of the model.

Algorithms to fit the model by MLE method, to calculate the EBLUP of small area means and to estimate its mean squared error are given in the paper. Further, an extensive simulation study is presented showing that if there are some outlying domains, i.e. the proposed model is true, and the standard nested error regression model is used, then a lack of precision is observed with respect to the proposed model. On the other hand, if the standard model is true and the proposed more complicated model is used, then the decrease of precision is negligible. An application using the above described Canary Islands 2013 data and estimating the domain total of unemployed people is also given. This application shows that the best model is not necessarily the model with only random or with only fixed effects, but somewhere in between, i.e. a model including both types of effects.

**II)** The nested error model assigns regression lines which are parallel with varying intercept to domains. Further flexibility in modeling may be achieved by using models allowing different slopes of the regression lines in different domains. For instance, it is conceivable that in some provinces the effect of the unemployment status is steeper than in others and that this steepness reflects policies, strategies and market conditions that differentiate provinces. One possibility to reflect these differences is the use of random regression coefficient models which allow the

coefficients of auxiliary variables to vary across domains. This type of models were proposed for the first time in Moura nad Holt (1999). In the paper

Hobza, T., Morales, D. (2011). Small Area Estimation of Poverty Proportions under Random Regression Coefficient Models. *Modern Mathematical Tools and Techniques in Capturing Complexity - Understanding Complex Systems*, Springer Berlin, pp. 315 – 328,

presented in Section 5.4 on page 95, we develop the idea and present an application to estimation of poverty proportions in Spanish provinces. The employed random regression coefficient model has the form

$$y_{dj} = \sum_{k=0}^{p} \beta_k x_{kdj} + \sum_{k=0}^{p} u_{kd} x_{kdj} + e_{dj}, \quad d = 1, \ldots, D,\ j = 1, \ldots, n_d, \tag{3.7}$$

where $y_{dj}$ is the $j$th observation from area $d$, $x_{kdj}$ are auxiliary variables and $\beta_k$ are unknown regression parameters. Further, random regression coefficients $u_{kd} \overset{iid}{\sim} N(0, \sigma_k^2)$ and random errors $e_{dj} \sim N(0, w_{dj}^{-1}\sigma_e^2)$ are independent, $d = 1, \ldots, D$, $j = 1 \ldots, n_d$, $k = 0, \ldots, p$. If $x_{0dj} = 1$ for any $d$ and $j$ then the model (3.7) contains a random intercept of the form $\beta_0 + u_{0d}$ for area $d$. The model variance and covariance parameters are $\sigma_e^2$, $\sigma_k^2$, $k = 0, \ldots, p$ and the heteroscedasticity weights $w_{dj}$'s are supposed to be known.

The paper presents EBLUP estimates based on the proposed model and gives a closed-formula procedure to estimate the mean squared error of the EBLUP. Behaviour of the obtained EBLUP is compared by simulation experiment with the behaviour of the EBLUP based on the classical nested error model. In the practical application we use data from the 2006 Spanish Living Conditions Survey (SLCS) with the goal of estimating poverty proportions in the 52 Spanish provinces. Because poverty variable is dichotomic at the individual level, the sample data from SLCS is previously aggregated to the level of census sections. As auxiliary variables nationality and employment status are used. From the results we can conclude that the proposed EBLUP behaves more smoothly and gives the best results with respect to the mean squared error when compared to classical EBLUP or direct estimators.

**III)** In addition to the variability of the regression slopes across domains allowed in the previous model (3.7), it is also possible to encounter situations where some correlation is detected between the random slopes and random intercept. In a more advanced paper,

Hobza, T. and Morales, D. (2013). Small area estimation under random regression coefficient models. *Journal of Statistical Computation and Simulation*, 83(11), pp. 2160-2177,

presented in Section 5.5 on page 110, we work with model of the form (3.7) but with additional assumptions

$$\mathrm{E}(u_{0d}\, u_{kd}) = \tau_k, \qquad d = 1, \ldots, D, \quad k = 1, \ldots, p$$

and

$$\mathrm{E}(u_{0d_1}\, u_{kd_2}) = 0, \qquad \text{if} \quad d_1 \neq d_2\,.$$

If we assume $x_{0dj} = 1$ for all $d$ and $j$ then the correlation between the random intercept $u_{0d}$ and the random part of the $k$-th regression parameter $u_{kd}$ is within area $d$ modelled by means of the covariance $\tau_k$. It is obvious that the model (3.7) is a special case of the present model for the choice $\tau_k = 0$, $k = 1, \ldots, p$.

Again, the formulas defining the EBLUP of domain mean and a closed-formula procedure to estimate its error are given in the paper under the assumption that the REML estimates of model parameters are used. We also propose a statistical test for deciding between models with and without correlations. Several simulation studies showing behaviour of the modified model and its advantages with respect to the model (3.7) without correlations and the standard nested

error model are presented. Finally, we illustrate the methodology and carry out an analysis of the 2006 SLCS data. Our target variables are the province means of the household normalized net annual income and the considered auxiliary variables are the secondary education and the employed labour status. Although it seems that the correlations $\tau_k$ do not play an important role in the analysed data, the application shows that procedures using the models with random intercepts produce some gain of precision with respect to direct estimates and EBLUP based on the standard nested error model.

# Chapter 4

# Models for discrete responses

The models presented in the previous chapters are suitable for continuous variables $y$, in fact a normal distribution of the responses $y$ was assumed. However, in many cases the observations are discrete or categorical. For example, the variable of interest $y$ may be binary and indicate if a person is unemployed or not or it may represent count of individuals with some property in a family, small company, hospital etc. In such cases the small area quantities of interest are usually proportions or counts, for instance proportion or total of unemployed persons in the area. To also cover this type of problems, an extension of linear models called *generalized linear models* (GLM) was proposed in McCullagh and Nelder (1989). The generalization is done in two directions. First, expectation of the random variable $y$ need not to be connected directly to a linear combination of some covariates (like in linear models) but it is associated with a linear function of some covariates through a link function. Second, the distribution of the variable $y$ need not be normal but it is supposed to be a member of so called *exponential family* of distributions. The exponential family covers a variety of distributions that include normal, binomial, Poisson and multinomial as special cases. The GLM may be expressed by the formula

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \qquad i = 1, \ldots, n,$$

where $\mu_i = \mathrm{E}(y_i)$, $y_i$ are independent random variables from exponential family of distributions, $\boldsymbol{x}_i$ are known vectors of auxiliary variables, $\boldsymbol{\beta}$ is vector of unknown parameters and $g$ is a known link function.

## 4.1 Generalized linear mixed models

Although GLM's are widely used in practical applications, they are usually not general enough to be used in SAE since they do not cover the case when the observations are dependent, a case often encountered in SAE problems. So another extension is needed. The idea is similar to that used in linear mixed models, namely to include random effects and define a *generalized linear mixed model* (GLMM).

Suppose that given a vector of random effects $\boldsymbol{u}$ the responses $y_1, \ldots, y_n$ are conditionally independent such that the conditional distribution of $y_i$ given $\boldsymbol{u}$ is a member of the exponential family with probability density function

$$f_i(y_i|\boldsymbol{u}) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi)\right\}, \tag{4.1}$$

where $b(\cdot), a_i(\cdot), c_i(\cdot,\cdot)$ are known functions, and $\phi$ is a dispersion parameter which may or may not be known. Using the notation $\mu_i = \mathrm{E}(y_i|\boldsymbol{u})$ for the conditional expectation of $y_i$ given $\boldsymbol{u}$,

which is associated with $\theta_i$, the formula defining the GLMM is

$$g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{u}, \qquad i = 1, \ldots, n, \tag{4.2}$$

where

- $\boldsymbol{x}_i^T = (x_{i1}, \ldots, x_{ip})$ and $\boldsymbol{z}_i^T = (z_{i1}, \ldots, z_{iq})$ are known vectors of auxiliary variables
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown parameters (fixed effects)
- $\boldsymbol{u} = (u_1, \ldots, u_q)^T$ is a vector of random effects
- $g(\cdot)$ is a monotonous, differentiable function called link function.

It is further assumed that the vector of random effects is normally distributed, $\boldsymbol{u} \sim N(0, \Sigma_u)$, where the covariance matrix $\Sigma_u$ may depend on a vector $\vartheta$ of unknown variance components.

First examples of GLMM were given in MacGibbon and Tomberlin (1989) and McCullagh and Nelder (1989). Since then, GLLM's have received considerable attention in various fields like biology, medical research and surveys. Without being exhaustive we can mention e.g. Breslow and Clayton (1993), Malec et al. (1997), Ghosh et al. (1998), Lahiri and Maiti (2002), Ghosh et al. (2009), Erciulescu and Fuler (2013), Militino et al. (2015) and Boubeta et al. (2016). Even though these models are very useful, they bring some difficulties concerning inference about the parameters. The reason is the form of the log-likelihood function under a general GLMM,

$$\log f(\boldsymbol{y}) = \log \int_{\mathbb{R}^q} f(\boldsymbol{y}|\boldsymbol{u})\, f(\boldsymbol{u})\, d\boldsymbol{u} = \log \int_{\mathbb{R}^q} \left( \prod_{i=1}^{n} f_i(y_i|\boldsymbol{u})\, f(\boldsymbol{u}) \right) d\boldsymbol{u},$$

which typically does not have a closed-form expression since it involves high-dimensional integral that cannot be evaluated analytically. Moreover, this integral is usually difficult to evaluate even numerically. For example, imagine that we have one random effect for each small area, there are $D = 50$ small areas, the overall sample size is $n = 1000$ and the variables $y_i$ have discrete distribution so that the density $f(y_i|\boldsymbol{u})$ is in fact probability function the values of which are less than one. Then, the dimension of the integral is 50 and moreover, the integrand involves a product of 1000 terms with each term less than one. Such a product is numerically zero so it is difficult to evaluate the integral with Monte Carlo method. To overcome these difficulties some approximation of the integral have to be used or non-likelihood-based inference must be taken into account. For the ease of exposition we will illustrate these aspects as well as methods for prediction of the target variable on the logistic mixed model which is the most common model of the form (4.2). Moreover, we restrict ourselves to the unit level models in this chapter.

## 4.2 Unit level logistic mixed model

In this section we assume that the variable of interest is binary and we introduce a unit level logistic mixed model. Again let $D$ denote the number of small areas or domains and $\boldsymbol{u} = (u_1, \ldots, u_D)^T$ be a vector of independent and $N(0, \sigma^2)$ distributed random effects. About the target variable $y_{dj}$, representing the $j$th sample observation from domain $d$, we assume that its conditional distribution is Bernoulli with parameter $p_{dj}$, i.e.

$$y_{dj}|_{u_d} \sim Be(p_{dj}), \qquad d = 1, \ldots, D, \quad j = 1, \ldots, n_d.$$

For the probability function of this distribution it holds

$$P(y_{dj}|\boldsymbol{u}) = P(y_{dj}|u_d) = p_{dj}^{y_{dj}}(1 - p_{dj})^{1-y_{dj}} \quad \text{for} \quad y_{dj} \in \{0, 1\} \tag{4.3}$$

and the conditional expectation of $y_{dj}$ is

$$\mathrm{E}(y_{dj}|u_d) = p_{dj}\,.$$

It is easy to show that the probability function (4.3) belongs to the exponential family (4.1) and that there is no unknown dispersion parameter $\phi$ in this case.

The basic unit level logistic mixed model uses logit function as the link function and can be written in the form

$$g(p_{dj}) = \log\left(\frac{p_{dj}}{1-p_{dj}}\right) = \boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d, \qquad d = 1,\ldots,D, \quad j = 1,\ldots,n_d. \tag{4.4}$$

The area specific random effects $u_d$ are used to explain the between area variability of the response variable which is not captured by the auxiliary variables. From the formula (4.4) it follows that the probability $p_{dj}$ that the variable $y_{dj}$ will take on value 1 is modelled as

$$p_{dj} = \frac{\exp\{\boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d\}}{1+\exp\{\boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d\}}\,. \tag{4.5}$$

In the following we consider the logistic mixed model (4.4), but the procedures are applicable also to other models belonging to the class of GLMM models.

### 4.2.1 Estimation of parameters of logistic mixed model

Let us now turn our attention to estimation of the parameters of logistic mixed model. There are two unknown parameters in the model (4.4), namely the vector of regression parameters $\boldsymbol{\beta}$ and the variance component $\sigma^2$ of the random effects.

First we illustrate the classical approach using maximum likelihood method for obtaining parameter estimates. Under the assumed simple correlation structure of the vector $\boldsymbol{u}$, notice that $\Sigma_u = \sigma^2 I_{D\times D}$, the log-likelihood function can be written in the form

$$\begin{aligned} l(\boldsymbol{\beta},\sigma^2;\boldsymbol{y}) &= \log P(\boldsymbol{y}) = \log\int_{\mathbb{R}^D} P(\boldsymbol{y}|\boldsymbol{u})f(\boldsymbol{u})\,d\boldsymbol{u} = \log\int_{\mathbb{R}^D}\left(\prod_{d=1}^{D} P(\boldsymbol{y}_d|u_d)f(u_d)\right)d\boldsymbol{u} \\ &= \log\prod_{d=1}^{D}\int_{\mathbb{R}} P(\boldsymbol{y}_d|u_d)f(u_d)\,du_d = \sum_{d=1}^{D}\log\int_{\mathbb{R}}\prod_{j=1}^{n_d} P(y_{dj}|u_d)f(u_d)\,du_d\,, \end{aligned} \tag{4.6}$$

where the notation $\boldsymbol{y} = (\boldsymbol{y}_1^T,\ldots,\boldsymbol{y}_D^T)$, $\boldsymbol{y}_d = (y_{d1},\ldots,y_{dn_d})^T$, $d = 1,\ldots,D$, is used and $f$ is the probability density function of the normal distribution $N(0,\sigma^2)$. So the dimension of the involved integral is one under the present model but the problem with product of many terms less than 1 remains. Substituting the probability function (4.3) and the probabilities (4.5) into the formula (4.6) the log-likelihood function can be expressed as

$$l(\boldsymbol{\beta},\sigma^2;\boldsymbol{y}) = -\frac{D}{2}\log(2\pi\sigma^2) + \sum_{d=1}^{D}\sum_{j=1}^{n_d} y_{dj}\boldsymbol{x}_{dj}^T\boldsymbol{\beta} + \sum_{d=1}^{D}\log\int_{\mathbb{R}}\exp\left\{h_d(u_d;\boldsymbol{\beta},\sigma^2)\right\}\,du_d,$$

where

$$h_d(u_d;\boldsymbol{\beta},\sigma^2) = \sum_{j=1}^{n_d}\left[y_{dj}u_d - \log\left(1+\exp\{\boldsymbol{x}_{dj}^T\boldsymbol{\beta}+u_d\}\right)\right] - \frac{u_d^2}{2\sigma^2}\,.$$

The integrals

$$\int_{\mathbb{R}}\exp\left\{h_d(u_d;\boldsymbol{\beta},\sigma^2)\right\}\,du_d \tag{4.7}$$

cannot be solved analytically and some approximation must be used.

One possibility is to use Laplace approximation of the integral which consists in second order Taylor expansion of the function $h_d(u_d; \boldsymbol{\beta}, \sigma^2)$ around the argument of its maxima $u_d^*$. The obtained expansion of $h_d$ is a quadratic function and the integral (4.7) may be then approximated using the known Gauss integral. This way we obtain Laplace approximation of the log-likelihood function and by its maximization we get the desired estimates.

Another way of approximation the log-likelihood provides method of penalized quasi-likelihood (PQL) which was proposed in the context of generalized mixed models in Breslow and Clayton (1993). PQL method gives an approximation of the log-likelihood function which has basically the same form as the Laplace approximation except for one missing term which is neglected. The estimation procedure is thus simplified and quicker than the Laplace method. It is known that the PQL does not provide consistent estimates but for small sample sizes its behaviour is comparable with Laplace approximation. For more details about these methods see e.g. Demidenko (2004).

The second approach to estimation of parameters of the logistic mixed model is the use of methods which are not based on likelihood function. One such method was proposed by Jiang (1998) and is called *method of simulated moments* (MSM). This method approximates the method of moments which would use for estimation of the parameters $\boldsymbol{\beta}$ and $\sigma^2$ the following system of nonlinear equations

$$\begin{aligned} 0 &= \sum_{d=1}^{D}\sum_{j=1}^{n_d} \mathrm{E}(y_{dj})\, x_{djk} - \sum_{d=1}^{D}\sum_{j=1}^{n_d} y_{dj} x_{djk}, \qquad k = 1, \ldots, p, \\ 0 &= \sum_{d=1}^{D} \mathrm{E}(y_{d.}^2) - \sum_{d=1}^{D} y_{d.}^2\,, \end{aligned}$$

where $y_{d.} = \sum_{j=1}^{n_d} y_{dj}$. As the expectations appearing in the equations cannot be explicitly evaluated, they are approximated by Monte Carlo simulation. The resulting set of equations is then solved numerically using the Newton-Raphson algorithm. Jiang (1998) proved that MSM gives consistent estimators of model parameters. A detailed description of this method is given in the Section 5.6.

### 4.2.2 Prediction of functions of fixed and random effects

Let us assume that the unit level logistic mixed model (4.4) - (4.3) holds not only for the sample but also for all units of a population $U$ with domain population sizes $N_1, \ldots, N_D$. Let us denote the sample $\boldsymbol{y}_s = (\boldsymbol{y}_{1s}^T, \ldots, \boldsymbol{y}_{Ds}^T)^T$, $\boldsymbol{y}_{ds} = (y_{d1}, \ldots y_{dn_d})^T$, $d = 1, \ldots, D$, and consider the problem of predicting a function of fixed and random effects,

$$\xi = \xi(\boldsymbol{\beta}, \boldsymbol{u})\,, \tag{4.8}$$

on the basis of the sample $\boldsymbol{y}_s$. Prediction of some function of mixed effects plays an important role in small area estimation. We now describe two methods developed in this context.

#### Empirical best predictors

Jiang and Lahiri (2001) introduced *best predictor* (BP) of the parameter $\xi$ which is defined by the formula

$$\widehat{\xi}^{bp} = \mathrm{E}(\xi|\boldsymbol{y}_s) = \mathrm{E}(\xi(\boldsymbol{\beta}, \boldsymbol{u})|\boldsymbol{y}_s) = \frac{\int_{\mathbb{R}^D} \xi(\boldsymbol{\beta}, \boldsymbol{u}) f(\boldsymbol{y}_s|\boldsymbol{u}) f(\boldsymbol{u}) d\boldsymbol{u}}{\int_{\mathbb{R}^D} f(\boldsymbol{y}_s|\boldsymbol{u}) f(\boldsymbol{u}) d\boldsymbol{u}}\,. \tag{4.9}$$

This predictor is "best" in the sense of minimization of the mean squared error in the class of predictors $\widehat{\xi}$ depending only on the sample $\boldsymbol{y}_s$, i.e.

$$\widehat{\xi}^{bp} \in \underset{\widehat{\xi}}{\operatorname{argmin}} \, \mathrm{E}\left(\widehat{\xi} - \xi\right)^2 .$$

Note that the BP depends on the sample $\boldsymbol{y}_s$ and the parameters $\boldsymbol{\beta}, \sigma^2$. Since the parameters are usually unknown, it is customary to replace them by appropriate estimators. The resulting predictor is called *empirical best predictor* (EBP).

For illustration, let us assume that we are interested in estimating the population means

$$\overline{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}, \qquad d = 1, \ldots, D\,.$$

These means may represent for example proportion of unemployed people in area $d$ and may be approximated by its conditional expectations, i.e. by the quantities

$$\overline{\mu}_d = \mathrm{E}\left(\overline{Y}_d | u_d\right) = \frac{1}{N_d} \sum_{j=1}^{N_d} p_{dj}, \qquad d = 1, \ldots, D. \tag{4.10}$$

Since the probabilities $p_{dj}$ are functions of the mixed effects (cf. formula (4.5)), we can use the general formula (4.9) to derive the best predictor

$$\widehat{p}_{dj}(\boldsymbol{\beta}, \sigma^2) = \mathrm{E}(p_{dj} | \boldsymbol{y}_s) = \frac{\int_{I\!R} \frac{\exp\{\boldsymbol{x}_{dj}^T \boldsymbol{\beta} + u_d\}}{1 + \exp\{\boldsymbol{x}_{dj}^T \boldsymbol{\beta} + u_d\}} P(\boldsymbol{y}_{ds} | u_d) f(u_d) du_d}{\int_{I\!R} P(\boldsymbol{y}_{ds} | u_d) f(u_d) du_d}\,. \tag{4.11}$$

The EBP of $p_{dj}$ and $\overline{\mu}_d$ are then $\widehat{p}_{dj}(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ and $\widehat{\overline{\mu}}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} \widehat{p}_{dj}(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$, respectively, and they can be approximated by Monte Carlo simulations.

### Plug-in predictor

González-Manteiga et al. (2007) proposed a simple (although not "best") predictor called *generalized EBLUP* (GEBLUP) or sometimes *plug-in* predictor in the literature. The plug-in predictor of the function $\xi(\boldsymbol{\beta}, \boldsymbol{u})$ given in (4.8) is obtained simply by substituting $\boldsymbol{\beta}$, $\boldsymbol{u}$ by parameter estimates $\widehat{\boldsymbol{\beta}}$ and random effect predictors $\widehat{\boldsymbol{u}}$, i.e.

$$\widehat{\xi}^{in} = \xi(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{u}})\,.$$

Similarly, the plug-in predictor of the probability $p_{dj}$ is

$$\widehat{p}_{dj}^{in} = \frac{\exp\{\boldsymbol{x}_{dj}^T \widehat{\boldsymbol{\beta}} + \widehat{u}_d\}}{1 + \exp\{\boldsymbol{x}_{dj}^T \widehat{\boldsymbol{\beta}} + \widehat{u}_d\}}\,.$$

Let us note that the prediction of the random effects $u_d$ is a separate task and can be solved in different ways. For example the PQL and Laplace methods for estimating parameters of the mixed model also provide predictors of the random effects $u_d$ which is not the case of the MSM. When MSM is used to estimate the parameters of the model, the EBP of the random effects $u_d$ may be used for construction of the plug-in predictor.

The main advantage of the plug-in predictors is their quite low computational demand in comparison with EBP's. For more details concerning the EBP and plug-in predictors and their comparison see the Section 5.6.

### 4.2.3 Mean squared error of predictors

In practise, it is desirable not only to compute the empirical predictors but also to asses their variation. Estimation of the mean squared error of empirical predictors under generalized linear mixed model is even more complex than under linear mixed model. The explicit expression of the exact MSE does not exist and only large sample approximations obtained under certain model assumptions are available. Jiang and Lahiri (2001) obtained an analytical approximation of the MSE of the EBP under logistic mixed model. Jiang (2003) extended their results to the class of generalized linear mixed models. González-Manteiga et al. (2007) gave an easy-to-apply closed formula estimator of the MSE of the GEBLUP when the GLMM is fitted by using the penalized maximum likelihood method. Molina et al. (2007) and López-Vizcaíno et al. (2013) extended the results of González-Manteiga et al. (2007) to multinomial-logit mixed models.

In Section 5.6 we give a detailed derivation of the approximation to the MSE of EBP of weighted sum of probabilities following the ideas of Jiang and Lahiri (2001) and Jiang (2003). A disadvantage of their method is that it requires analytical derivation of the formulas for each type of generalized linear mixed model and each function of mixed effects we want to predict. Moreover, these MSE estimates are computationally very demanding. For these reasons, resampling methods, which are applicable for estimating the MSE under more general model assumptions and are efficient and easy to implement, represent a good alternative.

Some resampling methods for estimating the MSE of empirical predictors can already be found in the literature. Jiang et al. (2002) introduced a jackknife methodology for MSE estimation. Pfeffermann and Tiller (2005) proposed parametric and nonparametric bootstrap methods for estimating the same quantity under state-space models. Hall and Maiti (2006a, 2006b) introduced parametric and matched-moment double-bootstrap algorithms, and González-Manteiga et al. (2007, 2008a, 2008b, 2010) applied bootstrap procedures to logistic and normal mixed models.

In Section 5.6 we give more detailed description of the bootstrap estimation of the MSE. Here we present just the basic idea of the parametric bootstrap technique for illustration. For the model (4.4), the procedure consists of the following steps:

1. Fit the model to the sample and calculate $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ and the EBP $\widehat{\xi}^{ebp}$ of the quantity of interest $\xi$.

2. Repeat $B$ times ($b = 1, \ldots, B$):

   a) Generate a new "bootstrap" population from the model (4.4) with parameters $\widehat{\boldsymbol{\theta}}$ and the same population sizes and covariates as the original population.

   b) Compute the true value $\xi^{true,b}$ of the quantity of interest $\xi$ for this population.

   c) From the bootstrap population select a bootstrap sample which has the same units as the original sample.

   d) For each bootstrap sample calculate the estimates $\widehat{\boldsymbol{\theta}}^{(b)}$ and the EBP $\widehat{\xi}^{ebp,b}$ of the quantity of interest $\xi$.

3. Output: the bootstrap MSE estimator

$$mse^*(\widehat{\xi}^{ebp}) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\xi}^{ebp,b} - \xi^{true,b}\right)^2 .$$

## 4.3 Contribution of the author to unit level logistic mixed models

**I)** A serious drawback of the best predictor of the form (4.11) is that if at least one of the auxiliary variables is continuous, the calculation of the EBP $\widehat{\overline{\mu}}_d$ requires the availability of census file with all the values of $\boldsymbol{x}_{dj}$ for all the $N_d$ units in area $d$. This is very restricting since in practice the full census records are rarely available. This is why in the paper

> Hobza, T., Morales, D. (2016). Empirical best prediction under unit level logit mixed models. *Journal of Official Statistics*, 32(3), pp. 661-692,

presented in Section 5.6 on page 129, we study the special case where the covariates are categorical and take a finite number of values. Let us assume that $\boldsymbol{x}_{dj} \in \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K\}$ for all $d$ and $j$ and define the probabilities

$$q_{dk} = \frac{\exp\{\boldsymbol{z}_k^T\boldsymbol{\beta} + u_d\}}{1 + \exp\{\boldsymbol{z}_k^T\boldsymbol{\beta} + u_d\}}, \qquad k = 1, \ldots, K.$$

Under this setup the target quantity $\overline{\mu}_d$ given in (4.10) can be rewritten as

$$\overline{\mu}_d = \frac{1}{N_d}\sum_{j=1}^{N_d} p_{dj} = \frac{1}{N_d}\sum_{k=1}^{K} N_{dk} q_{dk}, \qquad d = 1, \ldots, D, \tag{4.12}$$

where $N_{dk} = \#\{j \in U_d : \boldsymbol{x}_{dj} = \boldsymbol{z}_k\}$ is the population size of the covariate class $\boldsymbol{z}_k$ at the domain $d$. For calculating the EBP of $\overline{\mu}_d$ we thus do not need the full census file, it suffices to know the covariate class sizes $N_{dk}$ which can be obtained more easily from external sources.

In the above cited paper we consider the binomial-logit regression model

$$\log\left(\frac{p_{dj}}{1-p_{dj}}\right) = \boldsymbol{x}_{dj}^T\boldsymbol{\beta} + \phi v_d, \qquad d = 1, \ldots, D, \quad j = 1, \ldots, n_d, \tag{4.13}$$

for random variables $y_{dj}$ which conditional distribution is binomial,

$$y_{dj}|_{u_d} \sim Bin(m_{dj}, p_{dj}), \qquad d = 1, \ldots, D, \quad j = 1, \ldots, n_d,$$

where $m_{dj}$ are known size parameters. Let us note that the model formula (4.13) is only a reparametrization of the formula (4.4), actually, instead of random effects $u_d \sim N(0, \sigma^2)$ we use random effects $\phi v_d$ where $v_d \sim N(0, 1)$ and $\phi$ is a variance parameter corresponding to $\sigma$ in the previous notation.

The model parameters are estimated by the method of simulated moments and EBP and two plug-in estimators of the weighted sum of probabilities (4.12) are derived. For the EBP, we adapt the calculations given by Jiang and Lahiri (2001) and Jiang (2003) and we give two analytical estimators of the MSE approximation, without and with bias-correction term. Further, two parametric bootstrap estimators of the MSE are considered.

All the derived methods are compared via a computationally intensive simulation study. Jiang and Lahiri (2001) and Jiang (2003) studied the large sample properties of the EBPs and MSE estimators. However they did not carry out simulation experiments to empirically investigate the behavior of the EBPs and MSE estimators in the standard small area estimation setup, i.e. when the domain sample sizes are small. Results of our simulations show that the analytical approximations of the MSE work quite well but on the other side these estimators bring some implementational and computational difficulties which make their use in practical

applications almost impossible. For more details see Remark 5.1 in section 5.6. The bootstrap MSE estimators present quite good behaviour and may be considered as suitable alternative.

Finally, we present an application to estimation of poverty proportions in the counties of the region of Valencia in Spain. At the unit level, data are taken from the 2012 Spanish Living Conditions Survey (SLCS2012) and the target variables indicate whether individuals are under poverty line or not. As auxiliary variables we use the employment status of the individual (employed, unemployed, inactive, child) and the corresponding population sizes of covariate classes, $N_{dk}$, are obtained from the 2012 Spanish Labour Force Survey. The application shows some gain of precision obtained by the model-based EBP with respect to the direct design-based estimators.

**II)** Mixed models using temporal information are very useful because the recent past is generally very informative for the present. Temporal models thus borrow strength from the past for better estimation of the present. These models are employed in longitudinal studies with biological or medical data. In the context of small are estimation their use is more recent and one can find much more references dealing with the case of area level models. Time models at the unit level need more requirements than the area level models. In fact they need more elaborate software that has to be fed with data at the unit and at the aggregated level. In the paper

Hobza, T., Morales, D. (2016). Small area estimation of poverty proportions under a unit-level temporal binomial-logit mixed model. *TEST*, submitted pp. 1-20,

we deal with a unit level temporal binomial-logit mixed model with independent time effects for estimating poverty proportions and their changes between two consecutive years. Let us note that since the paper was not accepted at the time of finishing of this thesis it is not included in the Chapter 5. Here we give just its brief overview.

Let $D$ and $T$ be the number of small areas and time periods respectively. The model considers two independent sets of random effects such that $\{v_{1,d} : d = 1, \dots, D\}$ and $\{v_{2,dt} : d = 1, \dots, D,\ t = 1, \dots, T\}$ are independent and identically distributed (i.i.d.) $N(0, 1)$. These random effects are used to take into account the between-domains and the between-periods variability that is not explained by the auxiliary variables. The target variable $y_{dtj}$ represents the $j$th sample observation from domain $d$ at time period $t$ and its conditional distribution is supposed to be

$$y_{dtj}|_{v_{1,d},v_{2,dt}} \sim \text{Bin}(\nu_{dtj}, p_{dtj}), \quad d = 1, \dots, D,\ t = 1, \dots, T,\ j = 1, \dots, n_{dt}\,, \tag{4.14}$$

where $\nu_{dtj}$ is a known size parameter. The model is represented by the formula

$$\log \frac{p_{dtj}}{1 - p_{dtj}} = \boldsymbol{x}_{dtj}^T \boldsymbol{\beta} + \phi_1 v_{1,d} + \phi_{2t} v_{2,dt}, \quad d = 1, \dots, D,\ t = 1, \dots, T,\ j = 1, \dots, n_{dt}, \tag{4.15}$$

where $\phi_1 > 0$, $\phi_{2t} > 0$, $t = 1, \dots, T$, are variance parameters. We also consider two simpler models defined by the restrictions $\phi_{2t} = \phi_2$ or $\phi_{2t} = 0$ for all $t \in \{1, \dots, T\}$. Laplace approximation of the log-likelihood function of the model is derived and algorithm for estimation of the model parameters is given.

The main aim is to obtain empirical best predictors for the population averages

$$\overline{Y}_{dt} = \frac{1}{N_{dt}} \sum_{j=1}^{N_{dt}} y_{dtj}, \quad d = 1, \dots, D,\ t = 1, \dots, T.$$

The EBP of $\overline{Y}_{dt}$ can be expressed as the sum of two terms,

$$\hat{\overline{Y}}_{dt} = \frac{1}{N_{dt}} \Big[ \sum_{j=1}^{n_{dt}} y_{dtj} + \sum_{j \in U_{dt,r}} \hat{p}_{dtj} \Big],$$

where $U_{dt,r}$ denotes the non-sampled elements of the population in area $d$ at time $t$. The first term is a sum of observed values and the second term may again be treated as a sum of weighted probabilities if the covariates are categorical with finite possible values. Algorithms for calculation of the studied EBPs are given and their behaviour under the three assumed models is compared with the behaviour of the plug-in predictors by a simulation study. From the results it follows that the EBP and plug-in have similar behavior and the model without time effects gives the worst results.

We also revisit the application given in the previous part **I)** to data from the SLCS2012. This paper analyzes 2012-2013 data from the SLCS in the region of Valencia. The target of the application is the estimation of 2013 poverty proportions and 2013-2012 changes at county level. The results show that the EBP methodology is applicable to SAE real data problems and that the use of temporal information increases the precision of estimates.

## 4.4 Contribution of the author to unit level generalized linear models

In this section we describe several works which are themselves not directly connected to small area estimation but which could be applied to small area problems. These works deals with robust estimation and outlier detection in generalized linear models. The robust fitting of GLM is also of great importance for SAE. For example, plug-in estimators of domain proportions based on unit level logistic models and robust parameter estimators can be derived. Some robust approaches to SAE were already proposed in the literature. Let us mention e.g. the M-quantile regression assumed in Chambers and Tzavidis (2006) or the penalized spline regression used in Opsomer et al. (2008). The author's contribution to robust methods is the following.

**I)** In the first contribution,

Hobza T., Pardo L., Vajda, I. (2008). Robust Median Estimator in Logistic Regression. *Journal of Statistical Planning and Inference*, 138(12), pp. 3822-3840,

presented in Section 5.7 on page 162, we assume the logistic regression model

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}, \qquad i = 1,\dots,n,$$

where $p_i = \mathrm{E}(Y_i)$, $Y_1,\dots,Y_n$ are independent random variables with Bernoulli distribution $Be(p_i)$, $i = 1,\dots,n$, $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ are known regressors and $\boldsymbol{\beta} \in I\!R^d$ is an unknown vector of parameters. Notice that this model can be viewed as an unit level logistic model.

It is known that maximum likelihood estimates of the parameters $\boldsymbol{\beta}$ are sensitive to contamination of the observations $Y_1,\dots,Y_n$ by outliers or leverage points. For this reason, we propose in the above mentioned paper an $L_1$-estimator of parameters $\boldsymbol{\beta}$ which is based on median function and which is expected to be more robust than the MLE. Unfortunately, it is not possible to apply the median function of the Bernoulli observations $Y_1,\dots,Y_n$ directly since it is piecewise constant and it is thus not sensitive to small changes of the parameters $p_i$. So the basic idea of the paper is to assume a transformation, called *statistical smoothing*, of the discrete observations $Y_1,\dots,Y_n$. This transformation consists in adding independent and uniformly on $(0,1)$ distributed random variables $U_i$ to the observations $Y_i$, i.e. it considers the continuous data

$$Z_i = Y_i + U_i, \qquad i = 1,\dots,n,$$

where $U_i \overset{iid}{\sim} U(0,1)$. The *median estimator* $\widehat{\boldsymbol{\beta}}^{\mathrm{Me}}$ is then defined as

$$\widehat{\boldsymbol{\beta}}^{\mathrm{Me}} = \arg\min_{\boldsymbol{\beta}\in I\!R^d} \sum_{i=1}^{n} \left| Z_i - m\left(p(\boldsymbol{x}_i^T\boldsymbol{\beta})\right)\right| \tag{4.16}$$

where

$$p(\boldsymbol{x}_i^T\boldsymbol{\beta}) = \frac{\exp(x_i^T\boldsymbol{\beta})}{1+\exp(x_i^T\boldsymbol{\beta})}$$

and $m(p)$ is the median function

$$m\left(p\right) = F_p^{-1}\left(1/2\right) = \inf\left\{z \in I\!R : F_p\left(z\right) \geq 1/2\right\} \tag{4.17}$$

corresponding to the class of distribution functions $F_p$ of the random variables

$$Z = \mathrm{Be}\left(p\right) + \mathrm{U}(0,1)$$

when the parameter $p$ varies in the closed interval $[0,1]$. The median function (4.17) is strictly increasing in $p$ so the argument $m\left(p(\boldsymbol{x}^T\boldsymbol{\beta})\right)$ in (4.16) detects every change of the product $\boldsymbol{x}^T\boldsymbol{\beta}$.

Consistency and asymptotic normality of the median estimator are proved in the paper. Moreover a method of enhancing the median estimator is introduced. This method increases efficiency of the median estimator in some cases and consists in replacing the set of statistically smoothed data $Z_i = Y_i + U_i$, $1 \leq i \leq n$, by the expanded set obtained by considering for $k > 1$ the matrix of data

$$Z_{ij} = Y_i + U_{ij}, \qquad 1 \leq i \leq n, \qquad 1 \leq j \leq k, \tag{4.18}$$

where $U_{ij}$ are $U\left(0,1\right)$-distributed and mutually as well as on $Y_1, \ldots, Y_n$ independent random variables, and applying the median estimator to this expanded set. In other words the $k$-enhanced median estimator can be defined by

$$\widehat{\boldsymbol{\beta}}^{\mathrm{kMe}} = \arg\min_{\boldsymbol{\beta}\in I\!R^d} \frac{1}{k}\sum_{i=1}^{n}\sum_{j=1}^{k}\left|Y_i + U_{ij} - m\left(\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right)\right|. \tag{4.19}$$

Simulation studies are carried out to study the sensitivity of the median estimators to outlying and leverage points and to compare it with the sensitivity of some robust estimators previously introduced in the literature. The median estimators seem to be more robust for larger sample sizes and higher levels of contamination.

**II)** In the paper

> Hobza, T., Pardo, L. and Vajda, I. (2012). Robust median estimator for generalized linear models with binary responses. Kybernetika, 48(4), pp. 768-794,

presented in Section 5.8 on page 182, we generalize the results of the previous paper and we prove consistency and asymptotic normality of the median estimator also in other types of generalized linear models with binary responses. Namely, we deal with probit, log-log, complementary log-log, scobit and power logit models. Formulas for the asymptotic covariance matrix of the median estimator are derived under the above mentioned models. Results of simulation experiment studying the behaviour of the median estimator under the probit model are also reported.

**III)**

It seems that the idea of enhancing the median estimator can still be improved. If we let $k \to \infty$ in (4.19), we get the formula

$$\widehat{\boldsymbol{\beta}}^{\mathrm{MMe}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d+1}} \sum_{i=1}^{n}\int_0^1 \left|Y_i + u - m\left(\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right)\right| du \tag{4.20}$$

defining a deterministic estimate (i.e. it does not depend on any additionally generated random sample used for statistical smoothing), which would conceivably inherit the good properties of

the original estimate plus a smaller variance. This estimate is called *modified median estimator* and is treated in the paper

> Hobza, T., Martín, N., Pardo, L. (2016). A Wald-type test statistic based on robust modified median estimator in logistic regression models. *Journal of Statistical Computation and Simulation*, submitted pp. 1-22.

Since the paper was not accepted at the time of finishing of this thesis it is not included in the Chapter 5. Let us here just mention its principal ideas.

First, the asymptotic normality of the modified median estimator is proved. Further, based on the modified median estimator, we define a Wald-type test statistic for the problem of testing

$$H_0 : \boldsymbol{K}^T\boldsymbol{\beta} = \boldsymbol{m} \qquad \text{against} \qquad H_1 : \boldsymbol{K}^T\boldsymbol{\beta} \neq \boldsymbol{m}, \tag{4.21}$$

where $\boldsymbol{K}^T$ is any matrix of $r$ rows and $d$ columns and $\text{rank}(\boldsymbol{K}^T) = r$ and $\boldsymbol{m}$ is a vector of order $r$ of specified constants such that $\text{rank}(\boldsymbol{K}^T, \boldsymbol{m}) = r$. The statistic is given by the formula

$$W_n(\widehat{\boldsymbol{\beta}}^{\text{MMe}}) = n(\boldsymbol{K}^T\widehat{\boldsymbol{\beta}}^{\text{MMe}} - \boldsymbol{m})^T \left(\boldsymbol{K}^T\widehat{V}_n(\widehat{\boldsymbol{\beta}}^{\text{MMe}})\boldsymbol{K}\right)^{-1} (\boldsymbol{K}^T\widehat{\boldsymbol{\beta}}^{\text{MMe}} - \boldsymbol{m}), \tag{4.22}$$

where $\widehat{V}_n(\widehat{\boldsymbol{\beta}}^{\text{MMe}})$ is an estimator of the asymptotic covariance matrix of the modified median estimator. We show that under some regularity assumptions the asymptotic distribution of the Wald-type test statistics is a chi-square distribution with $r$ degrees of freedom and we derive approximation of the power function of the proposed test. An extensive simulation study is presented in order to analyze the efficiency as well as the robustness of the modified median estimator and Wald-type test based on it. The results show that the modified median estimator is much more efficient that the original median estimator and that the levels of the Wald-type tests are significantly more resistent to contamination of data by outliers than the levels of Bianco and Martínez (2009) robust test which was selected for comparison.

**IV)** In the last contribution included in this thesis,

> Pardo, M.C., Hobza, T. (2014). Outlier detection method in GEEs, *Biometrical Journal*, 56(5), pp. 838-850,

presented in Section 5.9 on page 210, we propose an outlier detection technique in the context of longitudinal data. The assumed model is similar to generalized linear model and can be written, in the notation of this thesis, as

$$g(\mu_{dj}) = \boldsymbol{x}_{dj}^T\boldsymbol{\beta}, \qquad d = 1, \ldots, D, \quad j = 1, \ldots, n_d,$$

where $g$ is a known link function, $\mu_{dj} = \text{E}(y_{dj})$ and the density of the response $y_{dj}$ is a member of the exponential family defined in (4.1). The main difference with respect to the classical general linear model is that the responses $y_{dj}$ need not be independent. More concretely, the elements of the vector $\boldsymbol{y}_d = (y_{d1}, \ldots, y_{dn_d})^T$ may be correlated but the vectors $\boldsymbol{y}_d$, $d = 1, \ldots, D$, are independent. In the SAE context it would mean that the observations in different areas are independent but the observations within an area are correlated. Notice that such correlation structure corresponds to the correlation structure of the mixed model (4.4). So the proposed method for detecting outliers could be used in the context of unit level logistic mixed models.

**Remark 6** Other interpretation of the assumed model is that $D$ is the total number of individuals in a study and for each individual $d$ we have a vector of repeated measurements $\boldsymbol{y}_d = (y_{d1}, \ldots, y_{dn_d})^T$ at $n_d$ time points. This interpretation is used in the above mentioned paper and corresponds to the typical definition of longitudinal data.

The parameters of the model are estimated by the method of generalized estimating equations (GEE) proposed by Liang and Zeger (1986). In order to describe the method for outlier detection let us denote $g(\boldsymbol{\mu}_d) = (g(\mu_{d1}), \ldots, g(\mu_{dn_d}))^T$, $d = 1, \ldots, D$, and $\boldsymbol{X}_d$ the matrix composed of the rows $\boldsymbol{x}_{dj}^T$, $j = 1, \ldots, n_d$. To identify an outlier in a designated area, say $i$, we propose to use the mean shift model

$$g(\boldsymbol{\mu}_d) = \begin{cases} \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\gamma}_i, & d = i; \\ \boldsymbol{X}_d\boldsymbol{\beta}, & d \neq i,\ d = 1, \ldots, D. \end{cases}$$

To test that area $i$ is an outlier (or contains an outlier) is equivalent to test the hypothesis

$$H_0 : \boldsymbol{\gamma}_i = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\gamma}_i \neq \mathbf{0}\,. \tag{4.23}$$

If the null hypothesis is rejected, the $i$-th area or its element will be highlighted as an outlier. Let us note that the choice $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{in_i})^T$ leads to the test whether the area $i$ as a whole is an outlier, whereas the choice $\boldsymbol{\gamma}_i = (0, \ldots, \gamma_{ij}, \ldots, 0)^T$ leads to the test whether the $j$-th observation in area $i$ is an outlier. For testing the hypothesis (4.23) we use the "working" score test studied by Rotnitzky and Jewell (1990).

The paper presents results of a simulation study which show that the proposed method correctly singles out the outlier when the data set have a known one. An application to real data set from a clinical trial is also given. The proposed approach detected the same outliers as the methods applied to this data set in the literature and moreover another one which was suspected to be an outlier by a visual scan of the data but which was not highlighted by any other method.

# Conclusion and possible future directions

This habilitation thesis deals with model-based methods for estimation of characteristics of areas or domains where the sample sizes are not large enough to provide precise direct estimates. Such areas are called small areas. Several modifications of linear models with random area effects for continuous data are proposed as well as some generalized linear mixed models for discrete data. For all the models the problems of estimation of the unknown parameters, best prediction of the characteristic of interest and estimation of the mean squared error of the prediction are treated. Results of the performed simulation experiments and applications to real data show that all the studied methods can be used in real data problems and moreover they provide a significant gain of precision with respect to classical methods or direct estimates. Further, methods proposed for robust estimation and outlier detection in the context of generalized linear models and their possible use in small area estimation problems are discussed.

In small area estimation, models with random effects for the areas introduce a correlation structure for the elements within the same area, but elements in different small areas are considered to be uncorrelated. However, it is known that socioeconomic characteristics of individuals in neighboring regions are usually more alike than those of individuals in distant regions. In statistical terms, this means that there is some kind of dependency relationship between individuals that are in neighboring regions. When this dependency is not completely captured by the auxi-liary variables in the model, it should be somehow incorporated in the correlation structure of the model. Not doing so may affect the performance of inferential procedures seriously (Cressie, 1993). Nevertheless, the introduction of a dependence structure among small areas entails a serious conceptual difference with respect to the traditional framework of independent small areas, in which the overall covariance matrix is block-diagonal. Thus, these models require new specific theoretical developments.

Some progress was already achieved in the frame of the basic Fay-Herriot model by Singh et al. (2005), Petrucci and Salvati (2006) and Pratesi and Salvati (2008), who considered an extension of the Fay-Herriot model by assuming that area effects follow a spatial autoregressive process of order 1, SAR(1). Bayesian spatial models have been considered by Moura and Migon (2002) and You and Zhou (2011).

Concerning possible extensions of the unit level temporal logistic mixed models considered in Section 4.3, a first step could be to consider the model (4.15) with autoregressive correlation structure of order 1, AR(1), for the random effects $\boldsymbol{v}_{2,d} = (v_{2,d1}, \ldots, v_{2,dT})$. By considering spatial dependence, a new extension of the model (4.15) might be proposed. An spatio-temporal model can be introduced by assuming that the area random effect $\boldsymbol{v}_1 = (v_{1,1}, \ldots, v_{1,D})$ has a SAR(1) distribution and the area-time random effects $\boldsymbol{v}_{2d}$, $d = 1, \ldots, D$, have independent AR(1) distributions. The derivation and implementations of SAE predictors based on such spatio-temporal models are tasks for future research.

# References


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

Bianco, A. M. and Martínez, E. (2009). Robust testing in the logistic regression model. *Computational Statistics and Data Analysis*, **53**, 4095–4105.

Boubeta M., Lombardía, M.J., and Morales, D. (2016). Empirical best prediction under area-level Poisson mixed models. *TEST*, **25**, 548-569.

Brackstone, G. J. (1987). Small area data: policy issues and technical challenges. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh eds.), 3-20. Wiley, New York.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

Chambers, R. and Tzavidis, N. (2006). M-quantile mod- els for small area estimation. *Biometrika*, **93**, 255–268.

Choudry, G.H. and Rao, J.N.K. (1989). Small area estimation using models that combine time series and cross sectional data. In: *Proceedings of Statistics Canada Symposium on Analysis of Data in Time* (Singh, A.C., Whitridge, P. Eds.), 67–74, Statistics Canada, Ottawa.

Cressie, N. (1993). *Statistics for spatial data*. John Wiley, New York.

Cressie, N. and Lahiri, S.N. (1993). The asymptotic distribution of RMLE estimators. *Journal of Multivariate Analysis*, **45**, 217–233.

Das, K., Jiang, J. and Rao, J. N. K. (2004). Mean Squared Error of Empirical Predictor. *The Annals of Statistics*, **32**, 818-840.

Datta, G.S. and Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *The Annals of Statistics*, **19**, 1748-1770.

Datta, G.S. and Lahiri, P. (2000). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems. *Statistica Sinica*, **10**, 613-627.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. John Wiley, New York.

Erciulescu A.L., Fuller W.A. (2013). Small Area Prediction of the Mean of a Binomial Random Variable. *JSM Proceedings. Survey Research Methods Section*. Alexandria, VA: American Statistical Association, 855-863.

Fay, R.E. and 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.

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

Ghosh, M., Kim, D., Sinha, K., Maiti, T., Katzoff, M., and Parsons, V.L. (2009). Hierarchical and Empirical Bayes small domain estimation and proportion of persons without health insurance for minority subpopulations. *Survey Methodology*, **35**, 53-66.

Ghosh, M., Nangia, N., and Kim, D. (1996). Estimation of median income of four-person families: a Bayesian time series approach. *Journal of the American Statistical Association*, **91**, 1423–1431.

Ghosh, M., Natarajan, K., Stroud, T.W.F., and Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, **93**, 273-282.

Ghosh, M. and Rao, J. (1994). Small area estimation: An appraisal. *Statistical Science*, **9**, 55-93.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D., and Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics and Data Analysis*, **51**, 2720-2733.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D. and Santamaría, L. (2008a). Bootstrap mean squared error of small-area EBLUP. *Journal of Statistical Computation and Simulation*, **78**, 443-462.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D., and Santamaría, L. (2008b). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Computational Statistics and Data Analysis*, **52**, 5242-5252.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D. and Santamaría, L. (2010). Small area estimation under Fay–Herriot models with nonparametric estimation of heteroscedasticity. *Statistical Modelling*, **10**(2), 215-239.

Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics*, **34**(4), 1733–1750.

Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society, Series B*, **68**(2), 221–238.

Henderson, C.R. (1950). Estimation of Genetic Parameters (Abstract), *Annals of Mathematical Statistics*, **21**, 309-310.

Jiang, J. (1998). Consistent estimators in generalized linear models. *Journal of the American Statistical Association*, **93**, 720-729.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of statistical planning and inference*, **111**, 117-127.

Jiang, J. and Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, **53**, 217-243.

Jiang, J., Lahiri, P., and Wan, S. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics*, **30**, 1782-1810.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, **15**, 1-96.

Kackar, R.N. and Harville, D.A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics - Theory and Methods*, **10**, 1249-1261.

Kackar, R.N. and Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, **79**, 853-862.

Lahiri, P. and Maiti, T. (2002). Empirical Bayes estimation of relative risks in disease mapping. *Calcutta Statistical Association Bulletin*, **53**, 211-212.

Lehtonen, R. and Veijanen, A. (2009). Design-based methods of estimation for domains and small areas. In *Sample Surveys: Inference and Analysis* (D. Pfeffermann and C. R. Rao, eds.). *Handbook of Statistics* 29B, 219–249. North-Holland, Amsterdam.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

López-Vizcaíno, E., Lombardía, M.J., and Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, **13**(2), 153-178.

MacGibbon, B. and Tomberlin, T.J. (1989). Small area estimation of proportions via empirical Bayes techniques. *Survey Methodology*, **15**, 237-252.

Malec D., Sedransk, J., Moriarity, C., 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.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, **15**, 1-24.

McClullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition. Chapman and Hall, London.

Militino A.F., Ugarte, M.D., and Goicoa, T. (2015). Deriving small area estimates from information technology business surveys. *Journal of the Royal Statistical Society, series A*, **178**(4), 1051–1067.

Molina, I., Saei, A. and Lombardíia, M.J. (2007). Small area estimates of labour force participation under multinomial logit mixed model. *Journal of the Royal Statistical Society, series A*, **170**, 975–1000.

Moura, F.A.S. and Holt, D. (1999). Small area estimation using multilevel models. *Survey Methodology*, **25**(1), 73–80.

Moura, F.A.S., Migon, H.S. (2002). Bayesian Spatial Models for small area estimation of proportions. *Statistical Modelling*, **2**(3), 183-201.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., and Breidt, F. J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, series B*, **70**, 265–286.

Petrucci, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 169–182.

Pfeffermann, D. (2002). Small area estimation-new developments and directions. *International Statistical Review*, **70**, 125-143.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**(1), 1-134.

Pfefferman, D. and Tiller, R. (2005). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, **26**, 893–916.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163-171.

Pratesi, M. and Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, **17**, 113–141.

Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, **25**, 175-186.

Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley, New York.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*, 2nd Edition. John Wiley, New York.

Rao, J.N.K. and Yu, M. (1994). Small area estimation by combining time series and cross sectional data. *Canadian Journal of Statistics*, **22**, 511–528.

Rotnitzky, A. and Jewell, N. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, **77**, 485–497.

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

Searle, S.R., Casella, G., and McCullogh, C.E. (1992). *Variance Components*. John Wiley, New York.

Singh, B., Shukla, G., and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, **31**, 183–195.

Stukel, D. (1991). *Small Area Estimation Under One and Two-Fold Nested Error Regression Model*. Ph.D. Thesis, Carleton University.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). *Finite Population Sampling and Inference. A Prediction Approach*, John Wiley, New York.

You, Y. and Zhou, Q.M. (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology*, **37**, 25-37.

# Chapter 5

# Relevant published articles of the author

## 5.1 A Fay-Herriot model with different random effect variances

What follows is a published article

Esteban, M.D., Herrador, M., Hobza, T., Morales, D. (2011). A Fay-Herriot model with different random effect variances. *Communications in Statistics – Theory and Methods*, 40(5), pp. 785-797.

A short description of this contribution can be found on page 25.

## 5.2 An Area-Level Model with Fixed or Random Domain Effects in Small Area Estimation Problems

What follows is a published article

Herrador, M., Esteban, M. D., Hobza, T., Morales, D. (2011). An Area-Level Model with Fixed or Random Domain Effects in Small Area Estimation Problems. *Modern Mathematical Tools and Techniques in Capturing Complexity - Understanding Complex Systems*, Springer Berlin, pp. 303 - 314.

A short description of this contribution can be found on page 26.

## 5.3 A modified nested-error regression model for small area estimation

What follows is a published article

Esteban, M.D., Herrador, M., Hobza, T., Morales, D. (2013). A modified nested-error regression model for small area estimation. *Statistics: A Journal of Theoretical and Applied Statistics*, 47(2), pp. 258-273.

A short description of this contribution can be found on page 30.

Pages 79 - 94 contain copyrighted material.

## 5.4 Small Area Estimation of Poverty Proportions under Random Regression Coefficient Models

What follows is a published article

Hobza, T., Morales, D. (2011). Small Area Estimation of Poverty Proportions under Random Regression Coefficient Models. *Modern Mathematical Tools and Techniques in Capturing Complexity - Understanding Complex Systems*, Springer Berlin, pp. 315 – 328.

A short description of this contribution can be found on page 31.

## 5.5 Small area estimation under random regression coefficient models

What follows is a published article

Hobza, T. and Morales, D. (2013). Small area estimation under random regression coefficient models. *Journal of Statistical Computation and Simulation*, 83(11), pp. 2160-2177.

A short description of this contribution can be found on page 31.

Pages 111 - 128 contain copyrighted material.

## 5.6 Empirical best prediction under unit-level logit mixed models

What follows is a published article

Hobza, T., Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. *Journal of Official Statistics*, 32(3), pp. 661-692.

A short description of this contribution can be found on page 39.

## 5.7 Robust Median Estimator in Logistic Regression

What follows is a published article

Hobza T., Pardo L., Vajda, I. (2008). Robust Median Estimator in Logistic Regression. *Journal of Statistical Planning and Inference*, 138(12), pp. 3822-3840.

A short description of this contribution can be found on page 41.

## 5.8 Robust median estimator for generalized linear models with binary responses

What follows is a published article

Hobza, T., Pardo, L. and Vajda, I. (2012). Robust median estimator for generalized linear models with binary responses. Kybernetika, 48(4), pp. 768-794.

A short description of this contribution can be found on page 42.

## 5.9 Outlier detection method in GEEs

What follows is a published article

Pardo, M.C., Hobza, T. (2014). Outlier detection method in GEEs, *Biometrical Journal*, 56(5), pp. 838-850.

A short description of this contribution can be found on page 43.

Pages 211 - 224 contain copyrighted material.

# Appendices

# Appendix A

# Curriculum vitae

# Curriculum Vitae

**Personal data:**

Name: **Tomáš Hobza**

Birth: **18. 1. 1975**

Nationality: **Czech**

Mail address: **FJFI ČVUT, Trojanova 13, CZ–12000 Praha 2**

Home address: **Halouny 68, 267 28 Svinaře**

Phone.: **+420 737 904 935**

E-mail: **hobza@fjfi.cvut.cz**

**Education:**

| | |
|---|---|
| 1993-1998 | Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, field of study: Mathematical Engineering, graduated on 8th June 1998 (**M.Sc. degree**); Thesis: *Modeling of Density Estimates* |
| 1998-2003 | Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Ph.D. student, field of study: Mathematical Engineering; Thesis defended on December 12, 2003 (**Ph.D. degree**); Thesis: *Asymptotics of Some Histogram-based Density Estimates* |

**Affiliation:**

| | |
|---|---|
| Sept. 1998 – Dec. 2011: | Academy of Sciences of the Czech Republic, half-time employee in the Institute of Information Theory and Automation (ÚTIA AV ČR), Department of Stochastic Informatics. Position: Young researcher |
| Since 2004: | Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering (ČVUT FJFI), Department of Mathematics. Present position: Assistant professor |
| **Teaching:** | Combinatorics and Probability; Mathematical Statistics; Applied Statistical Methods; Information Theory; Generalized Linear Models. Czech Technical University, since 2004 |
| **Research areas:** | Small Area Estimation: methods based on linear and generalized linear models, Divergence measures and their application in statistics |

**Participation in research grants:**

| | |
|---|---|
| 1998-1999 | participant of the grant EU - *Copernicus 579: Research of ATM networks,* coordinated by Igor Vajda, ÚTIA AV ČR |
| 1999-2001 | participant of the grant GACR n. 102/99/1137 *Estimation and optimization in telecommunication networks*, coordinated by Igor Vajda, ÚTIA AV ČR |
| 2000 | principle investigator of the CTU grant n. 300009704, *Optimization of binwidth in the generalized piecewise linear histogram.* |
| 2002-2004 | participant of the grant GACR n. 201/02/*1391, Asymptotic Properties of Information Contained in Quantized Observations*, coordinated by Igor Vajda, ÚTIA AV ČR |
| 2004-2005 | participant of the grant AV ČR n. A1075403, *New results in testing the goodness-of-fit based on Pearson-type statistics,* coordinated by Igor Vajda, ÚTIA AV ČR |
| 2007-2009 | participant of the grant GA ČR n. 102/07/1131, *Theoretical information in stochastic data and its empirical approximation and application in processes of detection and identification*, coordinated by Igor Vajda, ÚTIA AV ČR |

| | |
|---|---|
| 2010-2012 | participant of the grant GA ČR n. P202/10/0618, *Bregman distances, divergence of distributions, information retrieval, optimal decisions, machine learning*, coordinated by Igor Vajda, ÚTIA AV ČR |
| 2010-2012 | participant of the grant MTM2009-06997, *Fitting marginal models for longitudinal data*, coordinated by María del Carmen Pardo, UCM, Spain |
| 2015-present | participant of the grant GA ČR n. P103/15/15049S, *Detection of stochastic universalities in non-equilibrium states of socio-physical systems by means of Random Matrix Theory*, coordinated by Milan Krbálek, ČVUT FJFI. |

**Long-term stays abroad:**

| | |
|---|---|
| 2001: | 6 months research stay at the Miguel Hernández University, Elche, Spain (in the frame of the EU Socrates-Erasmus program) |
| 2003: | 1 month research stay at the Miguel Hernández University, Elche, Spain (in the frame of the EU Socrates-Erasmus program) |
| 2007: | 3 months research stay at the Miguel Hernández University, Elche, Spain (in the frame of the EU Socrates-Erasmus program) |
| 2011: | 1 month research stay at the Miguel Hernández University, Elche, Spain (in the frame of the EU Socrates-Erasmus program) |

**Other:**

- 2007 – 2010, 2010 – 2013, 2013 – 2016 member of the Academic Senate of the Faculty of Nuclear Sciences and Physical Engineering, since May 2014 chair of the Academic Senate.
- Language skills: English (postgraduate exam), Spanish (advanced), German (postgraduate exam)
- Scientometric data *(as of Jan 3, 2017);* citations (WoS, Scopus,…): 24 (14 without self-citation);

**Publication list** separately.

Prague, January 2017

Tomáš Hobza

# Appendix B

# List of publications

## Articles in impacted journals

1. Hobza, T., Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. Journal of Official Statistics, 32(3), pp. 661-692.

2. Krbálek, M., Hobza, T. (2016). Inner structure of vehicular ensembles and random matrix theory. Physics Letters A, 380(21), pp. 1839-1847.

3. Pardo, M.C., Hobza, T. (2014). Outlier detection method in GEEs, Biometrical Journal, Vol. 56(5), pp. 838-850.

4. Hobza, T., Morales, D., Pardo, L. (2014). Divergence-based tests of homogeneity for spatial data. Statistical Papers, 55(4), pp. 1059-1077.

5. Hobza, T. and Morales, D. (2013). Small area estimation under random regression coefficient models. Journal of Statistical Computation and Simulation, 83(11), pp. 2160-2177.

6. Esteban, M.D., Herrador, M., Hobza, T., Morales, D. (2013). A modified nested-error regression model for small area estimation. Statistics: A Journal of Theoretical and Applied Statistics, 47(2), pp. 258-273.

7. Hobza, T., Pardo, L. and Vajda, I. (2012). Robust median estimator for generalized linear models with binary responses. Kybernetika, 48(4), pp. 768-794.

8. Esteban, M.D., Herrador, M., Hobza, T., Morales, D. (2011). A Fay-Herriot model with different random effect variances. Communications in Statistics – Theory and Methods, 40(5), pp. 785-797.

9. Hobza T., Morales D., Pardo L. (2009). Rényi statistics for testing equality of autocorrelation coefficients. Statistical Methodology, 6(45), pp. 424-436.

10. Hobza T., Molina, I., Morales D. (2009). Multi-sample Rényi test statistics. Brazilian Journal of Probability and Statistics, 23(2), pp. 196-215.

11. Hobza T., Pardo L., Vajda, I. (2008). Robust Median Estimator in Logistic Regression. Journal of Statistical Planning and Inference 138(12), pp. 3822-3840.

12. Esteban, M.D., Hobza, T., Morales, D., Marhuenda, Y. (2008). Divergence-based tests for model diagnostic. Statistics and Probability Letters, 78(13), pp. 1702-1710.

13. Hobza, T., Molina, I., Vajda, I. (2005). On convergence of Fisher Informations in continuous models with quantized observation spaces. TEST, 14(1), pp. 151-179.

14. Hobza, T., Molina, I., Morales, D. (2003). Likelihood divergence statistics for testing hypothesis in familial data. Communications in Statistics – Theory and Methods, 32(2), pp. 415-434.

15. Berlinet, A., Hobza, T., Vajda, I. (2002). Generalized piecewise linear histograms. Statistica Neerlandica, 56(3), pp. 301-313.

16. Hobza, T., Vajda, I. (2001). On the Newcomb-Benford law in models of statistical data. Revista Matematica Complutense, 14(2), pp. 1-13.

## Reviewed articles in books

1. Herrador, M., Esteban, M. D., Hobza, T., Morales, D. (2011). An Area-Level Model with Fixed or Random Domain Effects in Small Area Estimation Problems. Modern Mathematical Tools and Techniques in Capturing Complexity - Understanding Complex Systems, Springer Berlin, pp. 303 - 314.

2. Hobza, T., Morales, D. (2011). Small Area Estimation of Poverty Proportions under Random Regression Coefficient Models. Modern Mathematical Tools and Techniques in Capturing Complexity - Understanding Complex Systems, Springer Berlin, pp. 315 – 328.

## Articles in international reviewed journals

1. Berlinet, A., Hobza, T., Vajda, I. (2002). Asymptotics for generalized piecewise linear histograms. Annals de l'Institute de Statistique de l'Université de Paris, 46(3), pp. 3-19.

## Other publications

1. Hobza, T. (2003). Asymptotics of some histogram-based density estimates. PhD dissertation, Czech Technical University in Prague, Czech Republic.

## Research reports

1. Esteban M., Hobza T., Marhuenda Y., Morales D. (2007). Divergence based statistics for model diagnostics. Research report n. I-2007-22, Centro de Investigación Operativa, Elche.

2. Hobza T., Morales D., Pardo L. (2007). Testing equality of autocorrelation coefficients in multivariate normal models. Research report n. I-2007-15, Centro de Investigación Operativa, Elche.

3. Hobza T., Molina, I., Morales D. (2007). Rényi statistics for testing hypotheses with s samples. Research report n. I-2007-30, Centro de Investigación Operativa, Elche.

4. Hobza T., Vajda I., van der Meulen E.C. (2006). Consistent Estimation and Testing by Means of Disparity Statistics Based on m-spacings. Research report DAR - ÚTIA 2006/15. ÚTIA AV ČR, Prague.

5. Hobza, T., Pardo, L., Vajda, I. (2006). Robust Median Estimator in Logistic Regression. Research report DAR - ÚTIA 2006/31. ÚTIA AV ČR, Prague.

6. Hobza, T. (2005). On the Consistency in Divergence for a Class of Nonparametric Distribution Estimates. Research report DAR - ÚTIA 2005/23. ÚTIA AV ČR, Prague.

7. Esteban, M. D., Hobza, T., Morales, D., Marhuenda, Y. (2005). Divergence-Based Tests for Model Diagnostic. Research report DAR - ÚTIA 2005/32. ÚTIA AV ČR, Prague.

8. Hobza, T., Pardo, L., Vajda, I. (2005). Robust Median Estimators in General Logistic Regression. Research report DAR - ÚTIA 2005/40. ÚTIA AV ČR, Prague.

9. Hobza, T., Pardo, L. and Vajda, I. (2004). Median estimators of parameters of logistic regression in models with discrete or continuous responses. Research report n. 2124, ÚTIA AV ČR, Prague.

10. Hobza, T., Molina, I., Morales, D., Pardo, L., Pardo, M.C., Rivas, M.J., Santo, M.T. and Vajda, I. (2003). Divergence statistics for testing composite hypotheses. Research report n. I-2003-1, Centro de Investigación Operativa, Universidad Miguel Hernández, Elche, Spain.

11. Hobza, T., Molina, I. and Morales, D. (2001). Rényi statistics for testing hypotheses with s samples, with an application to familial data. Research report n. I-2001-31, Centro de Investigación Operativa, Universidad Miguel Hernández, Elche, Spain.

12. Hobza, T., Kůs, V.,Vajda, I., van der Meulen, E. C. and Vrbenský, K. (1998). Optimal partitions and dominating distributions for Barron density estimates. Research report n. 1928, ÚTIA AV ČR, Prague.