Małgorzata K. KRZCIUK

Small area estimation – model-based approach in economic research

Publishing House of the University of Economics
in Katowice

Małgorzata K. Krzciuk

Small area estimation – model-based approach in economic research

Katowice 2023

Scientific publication

Monograph based on the dissertation awarded the first prize in the Statistics Poland President's Competition for the best doctoral thesis about statistics in 2022

Reviewer
Jacek Białek

Publishing Editor
Patrycja Keller

Typesetting
Małgorzata K. Krzciuk

Cover design
Emilia Gumulak

Cover picture © mindscanner – Photogenica

ISBN 978-83-7875-886-0

doi.org/10.22367/uekat.9788378758860

Publishing House of the University of Economics in Katowice
ul. 1 Maja 50, 40-287 Katowice, tel.: +48 32 257-76-33
www.wydawnictwo.ue.katowice.pl, e-mail: wydawnictwo@ue.katowice.pl
Facebook: @wydawnictwouekatowice

Contents

Introduction

The growing importance of regions and regional policy – regional programmes, participation in European Union programmes, development of regional self-governance – also entails an increase in the importance of national databases with a very detailed territorial division – sources increasingly used by public statistics, such as PESEL, POLTAX, POMOST, and ZUS. This is also related to the growing demand for information at an increasingly lower level of aggregation, as well as the demand for methods that do not require large financial outlays, but make it possible to obtain accurate estimations of subpopulation characteristics quickly, without the need for a full survey. Small area estimation methods may be the answer to this demand, allowing estimation and prediction under conditions where classical estimation methods prove to be inefficient or too costly. They allow estimation even for very small sample sizes, and even when the sample size of a subpopulation is zero. The choice of the topic considered in this monograph is therefore related to the increasing demand for local cross-sectional analyses. Moreover, it is also due to the multitude of fields in which the methods of small area estimation have already found application, such as market analyses, regional policy, labour market and poverty analysis, agricultural economics, and economic aspects of health policy.

The subject of this research will be the use of one of the main approaches in survey sampling, besides randomised and model-assisted – the model-based approach, in small area estimation for economic data. Aforementioned approach allows inference from purposive and random samples. The problem considered was the prediction of subpopulation characteristics and the analysis of predictor properties when there are different correlation relationships between random variables. The analyses took into account longitudinal data from the Local Data Bank, the largest organised collection of information in Poland on the socio-economic situation, demographics, and state of the environment, enabling multidimensional regional and local statistical analyses.

The main theoretical and exploratory objective of this book is to propose methods for predicting subpopulation characteristics and to analyse the properties of the predictors, taking into account the correlation between the random variables. The practical objectives include:

- adapting the methods of small area estimation, a model-based approach, for economic data obtained in longitudinal surveys;
- proposing and using the author's overpopulation models belonging to the class of general linear mixed models;
- proposing and using original model verification methods;
- proposing and applying the author's methods of prediction and assessment of prediction accuracy of subpopulation characteristics for the proposed class of models;
- demonstrating the applicability of the proposed methods to real economic data – simulation studies conducted using the Monte Carlo method.

The implementation of the above objectives will serve to answer the following research questions:

- Which models, belonging to the class of general linear mixed models, make it possible to take into account the occurrence of correlational relationships between random variables for prediction based on economic data obtained in longitudinal surveys?
- How can the presence of a correlation between random effects be verified for the proposed class of models?
- What effect does the inclusion of the presence of a correlation between random effects have on the properties of the considered predictors of subpopulation characteristics?
- How will the use of prior period information affect the accuracy of the considered predictors compared to methods using single-period information?

The monograph consists of five chapters. Each of them begins with an introduction. The first chapter of book discusses the theoretical basis of small area estimation. It presents the main approaches in small area estimation, including basic definitions, and issues concerning their development. Also presented are their selected areas of application in research of an economic nature, with examples. However, the greatest emphasis was placed on the presentation of the model-based approach. The process of building superpopulation models and their classification were discussed more extensively. Special attention was paid to the class of linear mixed models. The chapter presents the author's proposals for some special cases of models of this class with correlated vectors of random effects with their applications in small area estimation and generalisations of selected predictors to the case of longitudinal data. The chapter also discusses the author's proposals for the possibility of using permutation tests and those based on the parametric bootstrap method in verifying the significance of the parameters of the proposed superpopulation models.

The second chapter deals with the issue of repeated surveys over time. It discusses the essence of statistical longitudinal studies, including the main reasons for interest in and development of this type of research. It presents a classification of repeated surveys in time, together with conducting schemes and examples of economic research conducted both in Poland and worldwide. The essence of panel studies, and studies with partial and complete rotation are discussed in more detail. The chapter emphasises the advantages and disadvantages of repeated surveys over time. It also discusses the benefits and limitations of this type of research in the context of analyses based on it.

Chapter three discusses the problem of prediction using BLUPs and EBLUPs, including those proposed by Henderson (1950) and Royall (1976). Particular attention was paid to these classes of predictors in terms of the classification of linear mixed models into type A and B models. The author's proposal for the use of EBLUP under the assumption of a linear mixed model with correlated random effects in small area estimation is also presented. The chapter also addresses the issue of possible modifications as well as properties of the EBLUP class. Modifications of known methods for estimating the mean squared prediction error allowing for the estimation of the accuracy of EBLUP taking into account the correlation between random effects vectors are proposed. This part of the paper also included a review of selected economic applications of the above predictors.

The fourth chapter is focused on the class of empirical best predictors and plug-in predictors. The author's proposals for the use of predictors belonging to these classes in prediction based on models with correlated random effects vectors are presented. The problem of evaluating the mean squared errors of the proposed predictors is also addressed. The chapter also presents selected examples of applications of the discussed predictors in economic research.

Chapter five provides a description of the actual data set considered in the following section. It also presents the assumptions and results of the simulation studies carried out. Each of the subchapters was focused on one of the analysis variants, each of which was carried out according to the model-based approach. The problem of predicting total values and medians in domains under the assumption of a linear mixed model taking into account the correlation between random effects was addressed. The properties of the three proposed predictors belonging to the BP, EBP and plug-in classes were simulation tested. A comparison was made with selected predictors, assuming no correlation between random effects and selected estimators. This chapter, like the others, concludes with a brief summary of the issues raised. In the monograph, the author used methods of mathematical statistics and multivariate statistical analysis, as well as computer simulation techniques. The simulation studies used self-written programs in the R language (R Core Team, 2022). Analyses were conducted using actual data from several periods.

Chapter 1

Theoretical foundations of small area estimation

This chapter discusses the theoretical foundations of small area estimation. In the following subchapters, the main approaches in small area estimation are presented, including issues concerning their development and applications. In addition, for the model-based approach, the process of building overpopulation models and their classification is discussed in more detail. Particular attention is paid to the class of linear mixed models with correlated and uncorrelated random effects.

1.1. Main approaches in small area estimation – basic definitions and notation

Small area estimation is a branch of statistics covering methods enabling inference about the characteristics under study in distinguished subpopulations (domains) on the basis of data obtained from a sample survey and additional information from, inter alia, censuses or registers. Important issues in small area estimation include making inferences based on a sample of small or even zero size in the domain under study and making the most of available additional information. Among the main approaches in small area estimation we can mention: randomised, model-based, and model-assisted.

Important concepts for all approaches in small area estimation are population and sample. A finite N-element population Ω is a set of N objects such that $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ with $N < \infty$. Population elements are identifiable when they can be uniquely numbered from 1 to N and each element corresponding to a given number is observable (Cassel et al., 1977, p. 4). The (unordered) sample s of n-elements is any subset of the set Ω with number of elements n (cf. Bracha, 1996, pp. 18–19). Tillé (2006) defines a sample as a column vector such that $s = (s_1, \ldots, s_k, \ldots, s_N)$, where s_k for samples without replacement takes the value 1 when the k-th element is in the sample and 0 otherwise. For randomised samples with replacement, s_k may take values greater than 1 when an element has been selected for sampling several times (Tillé, 2006, p. 8). An ordered sample is an ordered sequence of elements $\underline{s} = (k_1, k_2, \ldots, k_i, \ldots, k_n)$, where

$1 \leqslant k_i \leqslant N$ and $1 \leqslant i \leqslant n$ (the indices k_i need not be different) (cf. Bracha, 1996, p. 17). The effective sample size is the number of non-repeating sample elements. For unordered samples, the effective sample size is equal to the sample size. The set S of all samples of type s is called the sample space.

Another important concept is the sampling frame, which is an inventory of elements belonging to the population, or at least disjoint subsets of the population (clusters). It should be characterised by: completeness, timeliness and identifiability. It must therefore contain up--to-date information on all the units belonging to the population and allow the retrieval of each unit that has been included in the sample (cf. Bracha, 1996, pp. 26–27).

A domain, also referred to as a study domain, is a subset of the Ω population. A small domain, following Rao and Molina (2015), is a domain whose sample size is small, insufficient to obtain estimates of domain characteristics by direct methods, thus using information about the variable under study only from the domain under analysis, with sufficient accuracy. It should be noted that the division into domains can be distinguished not only on the basis of geographical or administrative division criteria, but also economic or socio-demographic criteria (Rao and Molina, 2015, p. 3).

A trait and a trait parameter in a population or domain are also important concepts. A trait, otherwise known as a variable, is a function defined on a set of Ω such that $Y\colon \Omega \to R$ (Bracha, 1996, p. 14). The estimated parameter of a trait denoted by θ, following Wywiał (2010), can be descriptive parameters of the structure, such as e.g. mean value, total value or standard deviation.

Also among the key concepts in small area estimation is the concept of statistics. A statistic $Z = z(M)$, in terms of the design-based approach, is a function defined on the space of a random variable M ($M = \{(i, y_i) : i \in S\}$), such that for each $s \in S$, the function $z(m)$ ($m = \{(i, y_i) : i \in s\}$) depends on a vector of values of the test variable $\mathbf{y}$ through y_i, where $i \in s$ (cf. Cassel et al., 1977, p. 20; Bracha, 1996, p. 35). Analogously to the design-based approach, it is also possible to define the concept of statistics in the model-based approach. In this case, it is the function $\hat{\theta}(M^*)$ ($M^* = (i, Y_i) : i \in S$), such that for any s, realisation of the random variable S, the function $\hat{\theta}$ depends on $Y_1, Y_2, \ldots, Y_N$ through Y_i, where $i \in s$ (cf. Cassel et al., 1977, p. 91).

Also linked to the term "statistics" are the concepts of estimator and predictor used in randomised and model-based approaches. In these approaches, an estimator and a predictor are statistics that allow the assessment of the parameter θ. It should be noted that the basic classification of estimators and predictors used in survey sampling and small area estimation literature allows them to be divided into two classes: direct and indirect. An indirect estimator or predictor uses information about the study variable from outside the analysed domain or period. This allows estimation or prediction even when the sample size of observations from a small area is zero and the use of direct methods is not possible.

Taking the type of information about the study variable used in the estimation process as a classification criterion, Schaible (1993) distinguishes three classes of indirect estimators:

- domain indirect estimators – using data from another small area, but not from other periods,
- indirect estimators of time – using information from other periods, but only for the domain under study,
- indirect estimators of time and domain – taking into account information for another domain, in periods other than the period under consideration.

This division will be important, in particular, when the data under consideration are longitudinal.

The remainder of this chapter will discuss in more detail all the approaches mentioned in this section that are considered in small area estimation. For each approach, the basic definitions and notation will be presented, as well as the selected estimators or predictors.

1.1.1. Design-based approach

The origins of the design-based approach in survey sampling, of which small area estimation is a branch, can be traced back to the late 19th century. One of its precursors, according to Balicki (1989), is Kiaer (1897). In this approach, the vector of values of the studied characteristic is treated as non-random. Thus, also the characteristic of interest, e.g., $\theta_d = \frac{1}{N_d}\sum_{i=1}^{N_d} y_{id}$, is non-random. The only source of randomness in this approach is the sampling design.

The sampling design is called the probability distribution $P(s)$, defined on the sample space S where for each sample $s \in S$, the conditions are satisfied (Cassel et al., 1977, p. 9): $P(s) \geqslant 0$ and $\sum_{s \in S} P(s) = 1$. A sampling scheme, however, is a mechanism for drawing units into a sample that enables the implementation of a sampling design (Cassel et al., 1977, p. 15). Also related to the notion of a sampling design is the sampling strategy, which, for the parameter θ, is an ordered pair $(\hat{\theta}, P(s))$, where $\hat{\theta}$ is the estimator of the parameter θ. It should be added, following Rao (1962), that for each sampling design, there is at least one sampling scheme implementing that design.

Another important concept in the design-based approach is the r-th inclusion probability $\pi_{k_1,\ldots,k_r}$ – probability of selecting population elements $k_1,\ldots,k_r$ into the sample s:

$$\pi_{k_1,\ldots,k_r} = \sum_{s \in A(k_1,\ldots,k_r)} P(s),$$

where $A(k_1,\ldots,k_r) = \{s : k_i \in s, \text{ for } i = 1,\ldots,r\}$ (cf. Cassel et al., 1977, p. 11). Thus, the first-order inclusion probability π_k is the probability of selecting for the sample s, the k-th element and the second-order inclusion probability – the k-th and l-th elements ($k \neq l$) (Särndal et al., 1992, pp. 30–31). Following Tillé (2006), the above definitions of inclusion probabilities can be written as: $\pi_k = Pr(S_k > 0)$ and $\pi_{kl} = Pr(S_k > 0 \wedge S_l > 0)$, where $\mathbf{S} = (S_1, S_2, \ldots, S_N)$ is a sample drawn from an N-element population (Tillé, 2006, p. 17). It should also be added that the probabilities of π_k and π_{kl} fulfill the following conditions (Bracha, 1996, p. 20): $0 \leqslant \pi_k \leqslant 1$ and $\max\{0, \pi_k + \pi_l - 1\} \leqslant \pi_{kl} \leqslant \min\{\pi_k, \pi_l\}$.

As has already been pointed out in this monograph, an estimator of the parameter $\theta \in \Theta$ is the statistic $\hat{\theta}$ with values belonging to the set Θ, the value of which represents an assessment of the parameter θ (cf. Wywiał, 2010, p. 35). Taking into account, moreover, the division into direct and indirect estimators discussed in this chapter, among the direct estimators, we can mention: the estimator proposed by Horvitz and Thompson (1952), the ratio estimator considered, among others, by Royall and Cumberland (1981) and Wu (1982), the regression estimator (Watson, 1937), and the POS estimators presented by Bracha (1996). Among the indirect estimators, nevertheless, we can distinguish a group of synthetic estimators, including, among others, the synthetic oridinary estimator, the synthetic ratio estimator, and the synthetic regression estimator (Domański and Pruska, 2001, pp. 42–43). When discussing the classification of estimators, one should also mention the composite estimators. Both direct and indirect estimators can be used to construct them (Rao and Molina, 2015, p. 57).

When discussing the concept of an estimator, it is important to mention its properties. The bias of the estimator resulting from the assumed sampling design (p-bias of estimator) is given by the following formula (Cassel et al., 1977, p. 26):

$$B_p(\hat{\theta}) = E_p(\hat{\theta}) - \theta. \tag{1.1}$$

It should be noted that if $B_p(\hat{\theta}) = 0$, the estimator is unbiased. The relative bias of the estimator has the form:

$$rB_p(\hat{\theta}) = \frac{B_p(\hat{\theta})}{|\theta|}. \tag{1.2}$$

The variance of the estimator in the design-based approach (p-variance of the estimator) is defined as (Cassel et al., 1977, p. 26):

$$D_p^2(\hat{\theta}) = E_p(\hat{\theta} - E_p(\hat{\theta}))^2. \tag{1.3}$$

Furthermore, the root of the expression (1.3), called p-standard error $D_p(\hat{\theta})$, is a measure of the estimation precision. The relative p-standard error is given by the formula (Żądło 2008, p. 24):

$$rD_p(\hat{\theta}) = \frac{D_p(\hat{\theta})}{|\theta|} 100\%. \tag{1.4}$$

An expression having the following form:

$$MSE_p(\hat{\theta}) = E_p(\hat{\theta} - \theta)^2 = D_p^2(\hat{\theta}) + B_p^2(\hat{\theta}) \tag{1.5}$$

is, however, the p-mean square error. The root of the above expression $RMSE_p(\hat{\theta})$ is a measure of the estimation accuracy (Cassel et al., 1977, p. 26). The relative root of the p-mean squared error is given by the formula:

$$rRMSE_p(\hat{\theta}) = \frac{RMSE_p(\hat{\theta})}{|\theta|} 100\%. \tag{1.6}$$

The lower the value of the root of (1.3) and (1.5), the higher the precision and accuracy of the estimate, respectively. The estimator of $\hat{\theta}$ is furthermore consistent if, in the case of sampling with a replacement for each ε, there is $\lim\limits_{n\to\infty} P\{|\hat{\theta} - \theta| > \varepsilon\} = 0$ and for sampling without replacement, $\hat{\theta} = \theta$ when $n = N$ (Särndal et al., 1992, pp. 166–168).

A selection of parameter estimators in the domain will be discussed below. For each estimator, a formula describing the p-variance will be presented together with the estimator.

The first of the estimators presented is the estimator proposed by Horvitz and Thompson (1952), also known as the expansion estimator. The starting point for presenting the form of this estimator for the total value and the mean in the domain will be the estimator for the total value in the population. If the condition is fulfilled that for any k, $\pi_k > 0$, the p-unbiased estimator for the total value in fixed population $\tilde{y} = \sum_{k=1}^{N} y_k)$ proposed by Horvitz and Thompson (1952) for any sampling design has the following form:

$$\hat{\theta}_{\Omega}^{HT} = \hat{\tilde{y}}_{\Omega}^{HT} = \sum_{k\in s} \frac{y_k}{\pi_k}, \tag{1.7}$$

where π_k is the first-order inclusion probability of the k-th element. It should be noted that the p-variance of the above estimator is given by the formula (Horvitz and Thompson, 1952, p. 670):

$$D_p^2(\hat{\tilde{y}}_{\Omega}^{HT}) = \sum_{k\in\Omega} \left(\frac{y_k}{\pi_k}\right)^2 \pi_k(1-\pi_k) + \sum_{k\in\Omega} \sum_{l\in\Omega, k\neq l} \frac{y_k y_l}{\pi_k \pi_l}(\pi_{kl} - \pi_k\pi_l), \tag{1.8}$$

when values π_k are greater than 0. When the effective sample size is fixed, the p-variance of the Horvitz–Thompson estimator is determined based on the following formula proposed by Yates and Grundy (1953):

$$D_{pYG}^2(\hat{\tilde{y}}_{\Omega}^{HT}) = \sum_{k\in s} \sum_{l\in s, k\neq l} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 (\pi_{kl} - \pi_k\pi_l). \tag{1.9}$$

If the condition $k \neq l$ is satisfied, the p-unbiased estimator (1.7) is given by the following formula (cf. Horvitz and Thompson, 1952, p. 670):

$$\hat{D}_p^2(\hat{\tilde{y}}_{\Omega}^{HT}) = \sum_{k\in s} \left(\frac{y_k}{\pi_k}\right)^2 (1-\pi_k) + \sum_{k\in s} \sum_{l\in s, k\neq l} \frac{y_k y_l}{\pi_k \pi_l} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}. \tag{1.10}$$

This statistic may, however, take negative values. In this case, if $\pi_{kl} - \pi_k \pi_l \geqslant 0$ (for $k = 1, \ldots, N$; $l = 1, \ldots, N$, $k \neq l$), we can use the estimator considered by Sen (1953) and Yates and Grundy (1953), which only takes non-negative values when:

$$\hat{D}^2_{pSYG}(\hat{\tilde{y}}^{HT}_{\Omega}) = \sum_{k \in s} \sum_{l \in s, k \neq l} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}}. \tag{1.11}$$

Furthermore, if the assumption that for every k and l $(k \neq l)$ $\pi_{kl} > 0$ is met, this estimator is p-unbiased. It can be seen that in order to calculate the p-variance estimators, the values of the inclusion probabilities of both first and second order are necessary. Assessment of the p-variance of the Horvitz–Thompson estimator is also possible based on first-order inclusion probability values only. The p-variance estimator considered by Matei and Tillé (2005) has the following form:

$$\hat{D}^2_{pH}(\hat{\tilde{y}}^{HT}_{\Omega}) = \sum_{k \in s} \sum_{l \in s} \frac{y_k y_l}{\pi_k \pi_l} D_{kl}, \tag{1.12}$$

where $D_{kl} = c_k - \frac{c_k^2}{\sum_{j \in s} c_j}$ if $k = l$, otherwise $D_{kl} = -\frac{c_k c_l}{\sum_{j \in s} c_j}$ and $c_k = \frac{n}{n-1}(1 - \pi_k)$. It should be added that the value of c_k was proposed in the work of Hájek (1981). Antal and Tillé (2014) note the high efficiency and low bias of the above estimator. In the case where the parameter to be estimated is a total value in the domain ($\tilde{y}_d$), the Horvitz–Thompson estimator will be given by the formula (1.7), where y_i is replaced by y_{id}, where $y_{id} = y_i$ if $i \in s_d$ and 0 otherwise. In this case, the p-variance will be given by the formula (1.8) and its assessment by the formulas (1.10), (1.11) and (1.12), where y_i is also replaced by y_{id}. If the characteristic of interest is the mean value in the domain ($\bar{y}_d$), the estimator has the following form:

$$\hat{\bar{y}}^{HT}_{\Omega_d} = \frac{1}{N_d} \hat{\tilde{y}}^{HT}_{\Omega_d}, \tag{1.13}$$

where $\hat{\tilde{y}}^{HT}_{\Omega_d}$ is given by the formula (1.7), where y_i is replaced by y_{id}. Between the p-variance of the HT estimator of the total value and the mean value in the domain, the following relation holds:

$$D^2_p(\hat{\bar{y}}^{HT}_{\Omega_d}) = \frac{1}{N_d^2} D^2(\hat{\tilde{y}}^{HT}_{\Omega_d}). \tag{1.14}$$

An analogous dependence also exists for the p-variance assessments.

Further estimators, which will be presented more extensively in this section, are synthetic estimators. These estimators, as reported by Gonzalez (1973), use direct estimators of a subpopulation larger than the domain, such as a stratum or the whole population. The synthetic ratio estimator of the total value in the domain for any sampling design is given by the formula (cf. Bracha, 1996, p. 260):

$$\hat{\theta}^{il-SYN} = \frac{\tilde{x}_{\Omega_d}}{\hat{\tilde{x}}^{HT}_{\Omega}} \hat{\tilde{y}}^{HT}_{\Omega} = \frac{\tilde{x}_{\Omega_d}}{\hat{\tilde{x}}_{\Omega}} \hat{\tilde{y}}^{il}_{\Omega}, \tag{1.15}$$

where $\tilde{x}_{\Omega_d}$ and $\tilde{x}_{\Omega}$ are the total value of the auxiliary variable in the domain and population, respectively, and $\hat{\tilde{x}}_{\Omega}^{HT}$ is the Horvitz–Thompson estimator of the total value of the auxiliary variable in the population. In addition, $\hat{\tilde{y}}_{\Omega}^{il}$ is the ratio estimator of the total value of the study variable in the population, determined by the following formula:

$$\hat{\tilde{y}}_{\Omega}^{il} = \frac{\hat{\tilde{y}}_{\Omega}^{HT}}{\hat{\tilde{x}}_{\Omega}^{HT}} \tilde{x}_{\Omega}. \tag{1.16}$$

The p-value of the above estimator is calculated as (Bracha, 1996, p. 260):

$$D_p^2 \left(\hat{\theta}^{il-SYN} \right) = D_p^2 \left(\frac{\tilde{x}_{\Omega_d}}{\hat{\tilde{x}}_{\Omega}} \hat{\tilde{y}}_{\Omega}^{il} \right) = \left(\frac{\tilde{x}_{\Omega_d}}{\hat{\tilde{x}}_{\Omega}} \right)^2 D_p^2 \left(\hat{\tilde{y}}_{\Omega}^{il} \right). \tag{1.17}$$

It should be added that the p-variance of the ratio estimator necessary to calculate the total value has approximately the following form:

$$D_p^2 \left(\hat{\tilde{y}}_{\Omega}^{il} \right) \approx \sum_{k=1}^{N} \sum_{l=1}^{N} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \left(\pi_{kl} - \pi_k \pi_l \right), \tag{1.18}$$

where $E_k = y_k - Bx_k$ and $B = \left(\sum_{k \in \Omega} y_k \right) / \left(\sum_{k \in \Omega} x_k \right)$. The p-variance of the synthetic ratio estimator can be estimated using the following statistic (Żądło, 2008, p. 69):

$$\hat{D}_p^2 \left(\hat{\theta}^{il-SYN} \right) = \left(\frac{\tilde{x}_{\Omega_d}}{\hat{\tilde{x}}_{\Omega}} \right)^2 \hat{D}_p^2 \left(\hat{\tilde{y}}_{\Omega}^{il} \right), \tag{1.19}$$

where:

$$\hat{D}_p^2 \left(\hat{\tilde{y}}_{\Omega}^{il} \right) \approx \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \right), \tag{1.20}$$

$e_k = y_k - bx_k$ and $b = \left(\sum_{k \in s} \frac{y_k}{\pi_k} \right) / \left(\sum_{k \in s} \frac{x_k}{\pi_k} \right)$.

The synthetic regression estimator of the total value in the domain for any sampling design is given by the following formula (cf. Bracha, 1996, p. 260):

$$\hat{\theta}^{reg-SYN} = N_d \left[\hat{\tilde{y}}_{\Omega}^{HT} + \hat{\beta} \left(\bar{x}_{\Omega_d} - \hat{\tilde{x}}_{\Omega}^{HT} \right) \right] = \frac{N_d}{N} \hat{\tilde{y}}_{\Omega}^{reg} + N_d \hat{\beta} \left(\bar{x}_{\Omega_d} - \bar{x}_{\Omega} \right), \tag{1.21}$$

where the regression estimator of the population total value can be written as:

$$\hat{\tilde{y}}_{\Omega}^{reg} = \hat{\tilde{y}}_{\Omega}^{HT} + \hat{\beta} \left(\tilde{x}_{\Omega} - \hat{\tilde{x}}_{\Omega}^{HT} \right), \tag{1.22}$$

and the estimator $\hat{\beta}$ is given by the formula:

$$\hat{\beta} = \frac{\sum_{k=1}^{n} \left(x_k - \hat{\tilde{x}}_{\Omega}^{HT} \right) \left(y_k - \hat{\tilde{y}}_{\Omega}^{HT} \right) \frac{1}{\pi_k}}{\sum_{k=1}^{n} \left(x_k - \hat{\tilde{x}}_{\Omega}^{HT} \right)^2 \frac{1}{\pi_k}}. \tag{1.23}$$

Assuming the approximation $\bar{x}_{\Omega_d} = \bar{x}_{\Omega}$ the p-variance of the above estimator is given by the formula (Bracha, 1996, p. 261):

$$D_p^2 \left(\hat{\theta}^{reg-SYN} \right) \approx D_p^2 \left(\frac{N_d}{N} \hat{\tilde{y}}_{\Omega}^{reg} \right) = \left(\frac{N_d}{N} \right)^2 D_p^2 \left(\hat{\tilde{y}}_{\Omega}^{reg} \right), \tag{1.24}$$

Where the p-variance of the regression estimator of the total value has the form:

$$D_p^2\left(\hat{\hat{y}}_{\Omega}^{reg}\right) \approx \sum_{k=1}^{N}\sum_{l=1}^{N} \frac{E_k}{\pi_k}\frac{E_l}{\pi_l}\left(\pi_{kl}-\pi_k\pi_l\right), \tag{1.25}$$

and: $E_k = y_i - B_2 x_k - B_1$, $B_2 = \frac{\sum_{k=1}^{N}(x_k-\bar{x}_{\Omega})(y_k-\bar{y}_{\Omega})}{\sum_{k=1}^{N}(x_k-\bar{x}_{\Omega})^2}$, and $B_1 = \bar{y}_{\Omega} - B_2\bar{x}_{\Omega}$. The p-variance estimator of the synthetic regression estimator can be determined based on statistics (Żądło, 2008, p. 76):

$$\hat{D}_p^2\left(\hat{\theta}^{reg-SYN}\right) = \left(\frac{N_d}{N}\right)^2 \hat{D}_p^2\left(\hat{\hat{y}}_{\Omega}^{reg}\right), \tag{1.26}$$

where the p-variance of the regression estimator is given by the formula:

$$D_p^2\left(\hat{\hat{y}}_{\Omega}^{reg}\right) \approx \sum_{k=1}^{n}\sum_{l=1}^{n} \frac{e_k}{\pi_k}\frac{e_l}{\pi_l}\left(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\right), \tag{1.27}$$

and, furthermore, $e_k = y_i - b_2 x_k - b_1$, $b_2 = \frac{\sum_{k=1}^{n}\left(x_k-\hat{\hat{x}}_{\Omega}^{HT}\right)\left(y_k-\hat{\hat{y}}_{\Omega}^{HT}\right)/\pi_k}{\sum_{k=1}^{n}\left(x_k-\hat{\hat{x}}_{\Omega}^{HT}\right)^2/\pi_k}$, $b_1 = \hat{\hat{y}}_{\Omega} - b_2\hat{\hat{x}}_{\Omega}$.

The last of the synthetic estimators presented in this paper is the synthetic oridinary estimator, which is given by the following form (cf. Bracha, 1996, pp. 259–260):

$$\hat{\theta}^{zw-SYN} = \frac{N_d}{\hat{N}}\hat{\hat{y}}_{\Omega}^{HT} = \frac{N_d}{N}\hat{\hat{y}}_{\Omega}^{il}, \tag{1.28}$$

where $\hat{\hat{y}}_{\Omega}^{HT}$ and $\hat{\hat{y}}_{\Omega}^{il}$ are determined from the formulas (1.15) and (1.7), respectively, and $\hat{N} =$ $= \sum_{i\in s}\frac{1}{\pi}$.

For the above estimator, the p-variance is given by the formula (Bracha, 1996, p. 259):

$$D_p^2\left(\hat{\theta}^{zw-SYN}\right) = \left(\frac{N_d}{N}\right)^2 \sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{y_k-\bar{y}_{\Omega}}{\pi_k}\right)\left(\frac{y_l-\bar{y}_{\Omega}}{\pi_l}\right)\left(\pi_{kl}-\pi_k\pi_l\right), \tag{1.29}$$

while its estimate can be obtained using the following statistics:

$$\hat{D}_p^2\left(\hat{\theta}^{zw-SYN}\right) = \left(\frac{N_d}{N}\right)^2 \sum_{k=1}^{n}\sum_{l=1}^{n}\left(\frac{y_k-\bar{y}_s}{\pi_k}\right)\left(\frac{y_l-\bar{y}_s}{\pi_l}\right)\left(\pi_{kl}-\pi_k\pi_l\right). \tag{1.30}$$

Above, it was shown how the p-variance of synthetic estimators can be estimated. The problem of estimating their p-mean squared error was considered by Rao and Molina (2015).

The third type of estimator relevant to the analyses conducted in this book is the composite estimator. The composite estimator can be written as a linear combination of the component estimators (Rao and Molina, 2015, p. 57):

$$\hat{\theta}^{COMP} = q\hat{\theta}_1 + (1-q)\hat{\theta}_2, \tag{1.31}$$

where $\hat{\theta}^{COMP}$ is the composite estimator of the parameter θ, $\hat{\theta}_1$ and $\hat{\theta}_2$ are the first- and second--order component estimators, respectively, and q is the assumed weight $(q \in [0,1])$. Calculating the p-mean square error of the above estimator, we use the formula:

$$\begin{aligned} MSE_p\left(\hat{\theta}^{COMP}\right) &= q^2 MSE_p\left(\hat{\theta}^{A}\right) + (1-q)^2 MSE_p\left(\hat{\theta}^{B}\right) \\ &\quad + 2q(1-q)E_p\left(\hat{\theta}^{A}-\theta\right)\left(\hat{\theta}^{B}-\theta\right). \end{aligned} \tag{1.32}$$

The optimal value of q, and therefore minimising (1.32), has the following form (cf. Rao and Molina, 2015, p. 57):

$$q_d^* = \frac{MSE_p\left(\hat{\theta}^B\right) - E_p\left(\hat{\theta}^A - \theta\right)\left(\hat{\theta}^B - \theta\right)}{MSE_p\left(\hat{\theta}^A\right) + MSE_p\left(\hat{\theta}^B\right) - 2E_p\left(\hat{\theta}^A - \theta\right)\left(\hat{\theta}^B - \theta\right)} \approx \frac{MSE_p\left(\hat{\theta}^B\right)}{MSE_p\left(\hat{\theta}^A\right) + MSE_p\left(\hat{\theta}^B\right)},$$

where the approximation is based on the assumption that $E_p\left(\hat{\theta}^A - \theta, \hat{\theta}^B - \theta\right)$ is small relative to the mean squared errors. Touching on the choice of the component estimators of the composite estimator, Longford (2005) proposes that they should be estimators of the parameter under study based on the same formula for the population and the domain. It should further be noted that compound estimators have the advantage of being able to reduce the p-error of the mean squared versus relative to its component estimators.

1.1.2. Model-based approach

The model-based approach has been developed in survey sampling since the late 1930s, and one of the first publications was the paper by Cochran (1939). In this approach, the vector of values of the trait under study is treated as a vector of realisations of random variables, so that the characteristic of interest, e.g. $\theta_d = \frac{1}{N_d}\sum_{i=1}^{N_d} Y_i$, is random.

An important concept for the model-based approach is the superpopulation model, which is a set of conditions defining the joint probability distribution of ξ of a vector of random variables $Y = [Y_1, Y_2, \ldots, Y_N]^T$ (cf. Cassel et al., 1977, pp. 81–82). The predictor, as mentioned above, is the statistic $\hat{\theta}(Y^*)$ used to predict the parameter θ (cf. Cassel et al., 1977, p. 91). Among both direct and indirect predictors, we can distinguish, among others, the simple predictor, the ratio predictor presented in the paper by Chaudhuri and Stenger (2005), and the multivariate regression predictor considered by Valliant et al. (2000). It is also possible to distinguish a class of predictors of the BLU (best linear unbiased predictors) and EBLU (empirical best linear unbiased predictors) types presented in Rao and Molina (2015).

Predictors belonging to the BLUP and EBLUP classes will be presented in more detail in Chapter 3 in the context of linear mixed models. Also relevant to the analyses presented in this book are the classes of best predictors (BP) and empirical best predictors (EBP), which will be discussed in Chapter 4.

The bias of the predictor resulting from the assumed superpopulation model (ξ-bias of the predictor) is given by the following formula (Cassel et al., 1977, p. 92):

$$B_\xi(\hat{\theta}) = E_\xi(\hat{\theta} - \theta). \tag{1.33}$$

In the case where $B_\xi(\hat{\theta}) = 0$, the predictor is ξ-unbiased. The relative ξ-bias is determined based on the formula:

$$rB_{\xi}(\hat{\theta}) = \frac{B_{\xi}(\hat{\theta})}{|E_{\xi}(\theta)|}. \tag{1.34}$$

The prediction error is given by the formula:

$$U_{\xi} = \hat{\theta} - \theta, \tag{1.35}$$

whereas its variance:

$$D^2_{\xi}(U) = Var_{\xi}(U) = E_{\xi}(U - E_{\xi}(U))^2. \tag{1.36}$$

It should be noted that the root of the above measure (prediction standard error) having the following form (Żądło, 2008, pp. 28–29):

$$D_{\xi}(\hat{\theta} - \theta) = \sqrt{Var_{\xi}(\hat{\theta} - \theta)} \tag{1.37}$$

is a measure of the prediction precision. The relative prediction standard error is given by the formula:

$$rD_{\xi}(\hat{\theta}) = \frac{D_{\xi}(U)}{|E_{\xi}(\theta)|} 100\%. \tag{1.38}$$

The root of the mean squared prediction error that has the form:

$$MSE_{\xi}(\hat{\theta}) = E_{\xi}(\hat{\theta} - \theta)^2 = Var_{\xi}(\hat{\theta} - \theta) + B^2_{\xi}(\hat{\theta}), \tag{1.39}$$

denoted by $RMSE_{\xi}(\hat{\theta})$ is a measure of the prediction accuracy (Żądło, 2015, p. 31). The following measure:

$$rRMSE_{\xi}(\hat{\theta}) = \frac{RMSE_{\xi}(\hat{\theta})}{|E_{\xi}(\theta)|} 100\% \tag{1.40}$$

is the relative prediction root mean square error.

The problem of the non-informativeness of the sampling design should also be mentioned. For a non-informative sampling design, following Cassel et al. (1977), we refer to the case when it does not depend on the study variable but only on a known matrix of auxiliary variables $\mathbf{X}$. Attention should also be paid to the consequences of the non-informativeness of the sampling design. Given two sources of randomness in sample selection – the sampling design $p(s)$ and the joint distribution ξ associated with the assumed superpopulation model – if the sampling design is non-informative, then any p-unbiased or ξ-unbiased predictor is also $p\xi$-unbiased. It is worth noting that the order of the operators can be changed. It follows that (cf. Cassel et al., 1977, pp. 90–94):

$$E_pE_{\xi}\left(\hat{\theta} - \theta\right) = E_{\xi}E_p\left(\hat{\theta} - \theta\right). \tag{1.41}$$

In the case of an non-informative sampling design, when the predictor minimises the mean squared error of the prediction for each s sample, the ξ-expected value of the p-mean squared error is also minimised (cf. Cassel et al., 1977, pp. 90–94):

$$\forall_s \, E_\xi \left(\hat{\theta} - \theta\right)^2 \to \min \quad \Rightarrow \quad E_\xi E_p \left(\hat{\theta} - \theta\right)^2 \to \min . \tag{1.42}$$

Thus, the search for a predictor that will minimise the ξ-expected value of the p-mean squared error can be restricted to predictors that minimise the mean squared error of the prediction for each sample s. The issue of informative and non-informative sampling designs has been raised by Nathan and Holt (1980), Raghunath (1990), Pfeffermann et al. (2001), Eideh and Nathan (2006), and Pedone and Romano (2011). It should be added that it is also possible to test the non-informativeness of the sampling design. This problem has been considered, in his study, by Pfeffermann (1993).

This subsection will also discuss a selection of BLUPs, both direct and indirect. For each predictor, the assumptions as well as the form of the mean squared error of the prediction will be presented, together with its estimator. It should be added that in the case of direct predictors, the parameter vector of the overpopulation model $\boldsymbol{\beta}$ may take different values in individual domains, in contrast to indirect predictors where $\boldsymbol{\beta}$ is assumed to be fixed for the whole population (or take different values in subpopulations larger than domains).

Let us introduce the notation. We consider longitudinal data from M periods. The population in period t $(t = 1, 2, \ldots, M)$ denoted by Ω_t with size N_t is divided into subpopulations (domains) Ω_{dt} $(d = 1, 2, \ldots, D,\ t = 1, 2, \ldots, M)$ with sizes N_{dt}, where $\bigcup_{d=1}^{D} \Omega_{dt} = \Omega_t$ and $\sum_{d=1}^{D} N_{dt} = = N_t$. The sample (random or non-random) in period t will be denoted by s_t and its size by n_t. Let $\Omega_{rt} = \Omega_t \setminus s_t$, $N_{rt} = \bar{\bar{\Omega}}_{rt}$, $s_{dt} = \Omega_{dt} \cup s_t$ and $N_{dt} = \bar{\bar{s}}_{dt}$, $\Omega_{rdt} = \Omega_{dt} \setminus s_d t$, and $N_{rdt} = \bar{\bar{\Omega}}_{rdt}$. Furthermore, $\Omega = \bigcup_{t=1}^{M} \Omega_t$, $N = \bar{\bar{\Omega}}$, $s = \bigcup_{t=1}^{M} s_t$, $n = \bar{\bar{s}}$, $\Omega_d = \bigcup_{t=1}^{M} \Omega_{dt}$, $N_d = \bar{\bar{\Omega}}_d$, $s_d = \bigcup_{t=1}^{M} s_{dt}$, $n_d = \bar{\bar{s}}_d$, $\Omega_r = \Omega \setminus s$, $N_r = \bar{\bar{\Omega}}_r$, $\Omega_{rd} = \Omega_d \setminus s_d$, and $N_{rd} = \bar{\bar{\Omega}}_{rd}$, where $\bar{\bar{\Omega}}$ is cardinality of Ω.

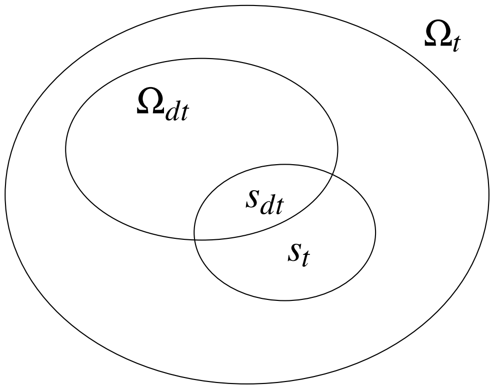


Figure 1.1. Notation in the t-th period of the longitudinal study

Source: Own elaboration.

It is worth noting that the above notation take into account the possibility that populations, subpopulations and the affiliation of population elements to subpopulations may change over time. Furthermore, they can be used for any type of longitudinal study. These notation is also presented in Figure 1.1.

The first predictor presented is a direct multivariate regression predictor. Let us make the following assumptions in this case (they are a generalisation of the longitudinal data model of the model considered by Żądło (2008)):

$$\begin{cases} E_{\xi}\left(\mathbf{Y}_{dt}\right) = \mathbf{X}_{dt}\boldsymbol{\beta}_{dt}, \\ D^2_{\xi}\left(\mathbf{Y}_{dt}\right) = \sigma^2_{dt}\mathbf{I}_{N_{dt}}, \end{cases} \tag{1.43}$$

where $\mathbf{Y}_{dt}$ is a vector of random variables of dimension $N_{dt} \times 1$, $\mathbf{X}_{dt}$ is a known matrix of auxiliary variables of dimension $N_{dt} \times p$, $\boldsymbol{\beta}_{dt}$ is a vector of unknown parameters of dimension $p \times 1$, σ^2_{dt} is an unknown parameter, and $\mathbf{I}_{N_{dt}}$ is a unit matrix of degree N_{dt}. Furthermore, independence of the random variables is assumed for observations from different domains and different periods. The BLUP of the total value in the d-th domain in period t is given by the formula (it is a direct generalisation to the case of longitudinal data of the predictor presented by Żądło (2008)):

$$\hat{\theta}^{reg}_{BLU} = \boldsymbol{\gamma}^T_s\mathbf{Y}_s + \boldsymbol{\gamma}^T_r\mathbf{X}_r\hat{\boldsymbol{\beta}}_{dt}, \tag{1.44}$$

where $\hat{\boldsymbol{\beta}}_{dt} = \left(\mathbf{X}^T_{s_{dt}}\mathbf{X}_{s_{dt}}\right)^{-1}\mathbf{X}^T_{s_{dt}}\mathbf{Y}_{s_{dt}}$. $\mathbf{Y}_s$ is a vector of random variables of dimension $n \times 1$, $\mathbf{Y}_{s_{dt}}$ is a vector of random variables of dimension $n_{dt} \times 1$, $\mathbf{X}_r$ is a known matrix of auxiliary variables of dimension $N_r \times p$, $X_{s_{dt}}$ is a known matrix of auxiliary variables of dimension $n_{dt} \times p$, and $\boldsymbol{\gamma}_s$ and $\boldsymbol{\gamma}_r$ are vectors of dimension $n \times 1$ and $N_r \times 1$, respectively, with elements equal to 1 for observations from the d-th domain in the t-th period and 0 otherwise. The mean squared error of the above predictor is given by the formula:

$$MSE_{\xi}\left(\hat{\theta}^{reg}_{BLU}\right) = Var_{\xi}\left(\hat{\theta}^{reg}_{BLU} - \theta\right) = g_1(\sigma^2_{dt}) + g_2(\sigma^2_{dt}), \tag{1.45}$$

where $g_1(\sigma^2_{dt}) = N_{rdt}$ and $g_2(\sigma^2_{dt}) = \sigma^2_{dt}\boldsymbol{\gamma}^T_r\mathbf{X}_r\left(\mathbf{X}^T_{s_{dt}}\mathbf{X}_{s_{dt}}\right)^{-1}\mathbf{X}^T_r\boldsymbol{\gamma}_r$. Calculation of the MSE score is possible in this case by replacing the estimator given by the formula in place of σ^2_{dt} in the above formula:

$$\hat{\sigma}^2_{dt} = \frac{1}{n_{dt} - p}\left(\mathbf{Y}_{s_{dt}} - \mathbf{X}_{s_{dt}}\hat{\boldsymbol{\beta}}_{dt}\right)^T\left(\mathbf{Y}_{s_{dt}} - \mathbf{X}_{s_{dt}}\hat{\boldsymbol{\beta}}_{dt}\right). \tag{1.46}$$

In the case of a direct ratio predictor, we also assume independence of the random variables and the model has the form (cf. Żądło, 2008, p. 98):

$$\begin{cases} E_{\xi}\left(Y_{idt}\right) = x_{idt}\beta_{dt}, \\ D^2_{\xi}\left(Y_{idt}\right) = \sigma^2_{dt}v(x_{idt}), \end{cases} \tag{1.47}$$

where $v(x_{idt})$ are the values of the known value function of the auxiliary variable. The BLUP is thus given by the formula (it is a direct generalisation to the case of longitudinal data of the predictor presented by Żądło (2008)):

$$\hat{\theta}^{il}_{BLU} = \sum_{i \in s_{dt}} Y_i + \hat{\beta}_{dt} \sum_{i \in \Omega_{rdt}} x_i, \tag{1.48}$$

where $\hat{\beta}_{dt}$ is calculated based on the following formula:

$$\hat{\beta}_{dt} = \left(\sum_{i \in s_{dt}} \frac{x_i^2}{v(x_i)} \right)^{-1} \sum_{i \in s_{dt}} \frac{x_i Y_i}{v(x_i)}. \tag{1.49}$$

The mean squared error of the above predictor has the following form:

$$MSE_{\xi}\left(\hat{\theta}^{il}_{BLU}\right) = Var_{\xi}\left(\hat{\theta}^{il}_{BLU} - \theta\right) = \sigma^2_{dt} \sum_{i \in \Omega_{rdt}} v(x_i) + \sigma^2_{dt} \left(\sum_{i \in s_{dt}} \frac{x_i^2}{v(x_i)} \right)^{-1}. \tag{1.50}$$

Replacing σ^2_{dt} in (1.50) with the estimator:

$$\hat{\sigma}^2_{dt} = \frac{1}{n_{dt} - 1} \sum_{i=1}^{n_{dt}} \frac{\left(Y_i - x_i \hat{\beta}_d\right)^2}{v(x_i)} \tag{1.51}$$

we will produce an estimator of the mean squared error of the presented predictor.

Let us generalise the considerations presented by Żądło (2008) and present below indirect regression and ratio predictors for longitudinal data. In the first case, let us assume a following superpopulation model of the form:

$$\begin{cases} E_{\xi}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \\ D^2_{\xi}(\mathbf{Y}) = \sigma^2 \mathbf{I}, \end{cases} \tag{1.52}$$

where independence of the random variables is assumed. The BLUP of the total value in the domain is given by the formula (1.44), however, $\hat{\boldsymbol{\beta}}_{dt}$ is replaced by: $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_s^T \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \mathbf{Y}_s$, where $\mathbf{X}_s$ is a known matrix of auxiliary variables of dimension $n \times p$, $\mathbf{Y}_s$ is a vector of random variables of dimension $n \times 1$, and n is understood as $n = \sum_{t=1}^{M} n_t$ (as declared before the introduction of the model (1.43)). We can determine the mean squared error of the indirect multivariate regression predictor based on the formula (1.45), where $g_1(\sigma^2_{dt})$ and $g_2(\sigma^2_{dt})$ are replaced by $g_1(\sigma^2) = N_{rdt}$ and $g_2(\sigma^2) = \sigma^2 \boldsymbol{\gamma}_r^T \mathbf{X}_r \left(\mathbf{X}_s^T \mathbf{X}_s\right)^{-1} \mathbf{X}_r^T \boldsymbol{\gamma}_r$. Calculation of the MSE score in this case is possible by replacing σ^2 with the estimator given by the formula:

$$\hat{\sigma}^2 = \frac{1}{n-p} \left(\mathbf{Y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}}\right)^T \left(\mathbf{Y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}}\right). \tag{1.53}$$

In the case of the indirect ratio predictor, independence of the random variables is also assumed and the model has the form:

$$\begin{cases} E_{\xi}(Y_{idt}) = x_{idt} \beta, \\ D^2_{\xi}(Y_{idt}) = \sigma^2 v(x_{idt}), \end{cases} \tag{1.54}$$

The BLUP in this case is given by the formula (1.48), where $\hat{\beta}_d$ is replaced by $\hat{\beta}$, determined by the following formula (cf. Żądło, 2008, p. 101):

$$\hat{\beta} = \left(\sum_{i \in s} \frac{x_i^2}{v(x_i)} \right)^{-1} \sum_{i \in s} \frac{x_i Y_i}{v(x_i)}. \tag{1.55}$$

The mean squared error of the above predictor is given by the formula (1.50), where σ^2_{dt} is replaced by σ^2 and the second sum is determined after the set s instead of s_{dt}. However, by replacing the estimator:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \frac{\left(Y_i - x_i \hat{\beta} \right)}{v(x_i)} \tag{1.56}$$

we obtain an estimator of the mean squared error of the indirect ratio estimator.

1.1.3. Model-assisted approach

The last approach discussed in small area estimation is the model-assisted approach. The polarisation between randomised and model-based approaches, according to Särndal (2010), occurred about 50 years ago. Both approaches found their proponents, and the differences between them and which statistics (estimators, predictors) are preferred under each approach became the focus of interest. One of the first papers to address the use of elements of both approaches was by Brewer et al. (1988). Already in its title, the authors ask the question: how reconcilable are model-based prediction and sampling design-based estimation? This question was also considered in the works of, among others: Smith (1994), Brewer (1995; 1999). One of the most important publications influencing the development of this approach is the work of Särndal et al. (1992). In both the randomised and model-based approaches, as Särndal (2010) states, there is a need to incorporate elements of these approaches into each other. For researchers following the design-based approach, the challenge is to explicitly define a model that can be used, e.g., in stratified variance modelling to determine sampling fractions from strata. For proponents of the model-based approach, this may include consideration of random sampling and selection of a sampling design with an assumed over-population model. It should be noted that within the methods belonging to the model-assisted approach, it is possible to distinguish those closer to the randomised or model-based approach. Statistics closer to the design-based approach include calibrated estimators (e.g. Deville and Särndal, 1992), while model-based – pseudo-empirical best linear predictors (e.g. Prasad and Rao, 1999).

In the following section, calibrated estimators for characteristics in the domain will be discussed in more detail, together with some modifications. The starting point for consideration is the calibrated estimator for the total value in the population presented in the work of Deville and Särndal (1992), which is given by the following formula:

$$\hat{\theta}_{\Omega}^{CAL} = \hat{\tilde{y}}_{\Omega}^{CAL} = \sum_{i \in s} w_{si} y_i, \tag{1.57}$$

where the weights w_{si} fulfil the conditions defined by the calibration equation:

$$\forall_{k \in \{1,2,\ldots,p\}} \sum_{i \in s} w_{si} x_{ik} = \sum_{i \in \Omega} x_{ik}. \tag{1.58}$$

Deville and Särndal (1992) also proposed that, in addition, the weights should take values as close as possible to the inverse of the first-order inclusion probabilities π_i. Thus, the second part of the task to designate the weights can be written as:

$$f_s(w_{si}, d_i, q_i) \to \text{min}, \tag{1.59}$$

where $f_s(w_{si}, d_i, q_i)$ is a function of the assumed distance of the weights w_{si} of the calibrated estimator and the weights $d_i = \frac{1}{\pi_i}$ of the Horvitz–Thompson estimator, and some additional weights are denoted by q_i. It should be noted that the above estimator (1.57) is asymptotically p-unbiased if there is a solution to the conditional minimisation task (1.59) under the condition (1.58). In addition, (1.58) is a ξ-unbiased condition under the assumption of a general linear model. Under an additional assumption:

$$f_s(w_{si}, d_i, q_i) = \sum_{i \in s} \frac{(w_{si} - d_i)^2}{d_i q_i}, \tag{1.60}$$

the solution to the above task will be the generalised regression estimator (GREG). This estimator has the following form:

$$\hat{\theta}^{GREG} = \sum_{i \in s} d_i y_i + \left(\sum_{i \in \Omega} \mathbf{x_i} - \sum_{i \in s} d_i \mathbf{x_i} \right)^T \hat{\mathbf{B}}, \tag{1.61}$$

where $\hat{\mathbf{B}} = \left(\sum_{i \in s} d_i q_i \mathbf{x_i} \mathbf{x_i}^T \right)^{-1} \sum_{i \in s} d_i q_i \mathbf{x_i} y_i$.

The asymptotic form of the p-variance of this estimator is given by the following formula:

$$\breve{D}^2 \left(\hat{\theta}^{GREG} \right) = \sum_{i \in \Omega} \sum_{j \in \Omega} (\pi_{ij} - \pi_i \pi_j) d_i E_i d_j E_j, \tag{1.62}$$

where $E_i = y_i - \mathbf{x}_i^T \mathbf{B}$ and $\mathbf{B} = \left(\sum_{i \in \Omega} q_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i \in \Omega} q_i \mathbf{x}_i y_i$. The following p-consistent estimator (Rao, 2003, p. 12) can be used to estimate its variance $\hat{\theta}^{GREG}$:

$$\hat{D}^2 \left(\hat{\theta}^{GREG} \right) = \sum_{j > i}^{n} \sum_{i}^{n} (\pi_i \pi_j - \pi_{ij}) \pi_{ij}^{-1} (d_i e_i - d_j e_j)^2. \tag{1.63}$$

Note that the above variance estimator is of the estimator form (1.11), where y_i is replaced by residuals of the form: $e_i = y_i - \mathbf{x}_i^T \hat{\mathbf{B}}$. Due to the underestimation of variance by the above estimator, the literature proposes using the following statistic, also being a p-consistent estimator, to estimate the variance of $\hat{\theta}^{GREG}$:

$$\hat{D}^2\left(\hat{\theta}^{GREG}\right) = \sum_{j>i}^{n}\sum_{i}^{n}\left(\pi_i\pi_j - \pi_{ij}\right)\pi_{ij}^{-1}\left(d_i g_{si} e_i - d_j g_{sj} e_j\right)^2, \tag{1.64}$$

where g_{si} is given by the following formula:

$$g_{si} = 1 + \left(\sum_{i\in\Omega}\mathbf{x}_i - \sum_{i\in s} d_i\mathbf{x}_i\right)^T \left(\sum_{i\in s} d_i q_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \mathbf{x}_i q_i. \tag{1.65}$$

When the parameter under consideration is the total value in the domain, the GREG estimator is given by the following formula (Rao, 2003, p. 17):

$$\hat{\theta}_d^{GREG} = \hat{\tilde{y}}_{\Omega_d}^{GREG} = \sum_{i\in s_d} w_{si} y_i. \tag{1.66}$$

The weights w_{si} have the form:

$$w_{si} = g_{si} d_i, \tag{1.67}$$

where g_{si} are given by the formula (1.65). It can therefore be seen that $\hat{\tilde{y}}_{\Omega_d}^{CAL} = \hat{\tilde{y}}_{\Omega}^{CAL}$ in (1.57) is replaced by y_{id}, where $y_{id} = y_i$ if $i \in s_d$ and 0 otherwise. It should be added that the use of the above estimator is not possible when $s_d = \emptyset$. Furthermore, this estimator does not require knowledge of the values of the auxiliary variables at the domain level. Even when the expected value of the abundance in the domain is small, it is approximately p-unbiased. The assessment of the variance of the GREG estimator of the total value in the domain can be determined based on the formula (1.63), however, the values of e_k should be replaced by (Rao, 2003, p. 17):

$$e_{id} = a_{id*} y_i - \mathbf{x}_i^T \hat{\mathbf{B}}_{d*}, \tag{1.68}$$

where a_{id*} takes the value 1 for $i \in \Omega_{d*}$ and zero otherwise, and $\hat{\mathbf{B}}_{d*}$ is given like in formula (1.61), where y_i is replaced by $a_{id*} y_k$. In the case of elements not belonging to Ω_{d*}, the residuals are of the form: $e_{id} = -\mathbf{x}_i^T \hat{\mathbf{B}}_{d*}$. The consequence of this can be, according to Rao (2003), inef-efficient variance estimation. If the characteristic of interest is the mean value in the domain, the GREG calibrated estimator is given by the following formula:

$$\hat{\bar{y}}_{\Omega_d}^{GREG} = \frac{1}{N_d}\hat{\tilde{y}}_{\Omega_d}^{GREG}, \tag{1.69}$$

where $\hat{\tilde{y}}_{\Omega_d}^{GREG}$ has the form (1.66). The variance estimator of the estimator in discussion is given by the formula:

$$\hat{D}^2(\hat{\bar{y}}_{\Omega_d}^{GREG}) = \frac{1}{N_d^2}\hat{D}^2(\hat{\tilde{y}}_{\Omega_d}^{GREG}), \tag{1.70}$$

where $\hat{D}^2(\hat{\tilde{y}}_{\Omega_d}^{GREG})$ is calculated on the basis of (1.63) and (1.68).

Särndal also considers some modification of the GREG estimator. The MGREG estimator for the total value in the domain is given by the following formula:

$$\hat{y}_{\Omega_d}^{MGREG} = \sum_{i \in s_d} d_i y_i + \left(\sum_{i \in \Omega_d} \mathbf{x}_i - \sum_{i \in s_d} d_i \mathbf{x}_i \right)^T \hat{\mathbf{B}} = \left(\sum_{i \in \Omega_d} \mathbf{x}_i \right)^T \hat{\mathbf{B}} + \sum_{i \in s_d} \frac{e_i}{\pi_i}, \tag{1.71}$$

where $\hat{\mathbf{B}} = \left(\sum_{i \in s} d_i q_i \mathbf{x_i} \mathbf{x_i}^T \right)^{-1} \sum_{i \in s} d_i q_i \mathbf{x}_i y_i$ and $e_i = y_i - \mathbf{x}_i^T \hat{\mathbf{B}}$. The first component of the above estimator, following Särndal and Hidiroglou (1989), is called synthetic, while the second is called correctional. It should be noted that this estimator has the following property:

$$\sum_{d=1}^{D} \hat{y}_{\Omega_d}^{MGREG} = \hat{y}^{GREG}. \tag{1.72}$$

In addition, Särndal and Hidiroglou (1989) suggest the following modification of this estimator, or more precisely of its correction component:

$$\hat{y}_{\Omega_d}^{MGREG} = \sum_{i \in \Omega_d} \mathbf{x}_i^T \hat{\mathbf{B}} + N_d \hat{N}_d^{-1} \sum_{i \in s_d} \frac{e_i}{\pi_i}, \tag{1.73}$$

where $\hat{N}_d = \sum_{i \in s_d} \frac{1}{\pi_i}$. For cases where the sample size is large, irrespective of the domain sample size, Rao and Molina (2015) propose to determine the variance of the above estimator from the formula (1.63) and thus the formula also used for GREG. In the above formula, e_i is replaced by:

$$e_{id}^{MGREG} = a_{id*}(y_i - \mathbf{x}_i^T \hat{\mathbf{B}}), \tag{1.74}$$

where a_{id*} takes the value of 1 for elements in the domain under study, and 0 in other cases. In the case of a zero sample size in the domain, the MGREG estimator is simplified to a synthetic estimator having the following form:

$$\hat{y}_{\Omega_d}^{MGREG} = \left(\sum_{k \in \Omega_d} \mathbf{x}_k \right)^T \hat{\mathbf{B}}.$$

In this case, given the formula (1.63), it is not possible to estimate its variance.

When the parameter being estimated is the mean value in the domain, the MGREG estimator is given by the formula:

$$\hat{\bar{y}}_{\Omega_d}^{MGREG} = \frac{1}{N_d} \hat{y}_{\Omega_d}^{MGREG}, \tag{1.75}$$

where $\hat{y}_{\Omega_d}^{MGREG}$ has the form (1.73). It should be added that the following dependence holds for the estimation of variance for the estimators of the data by the formulae (1.75) and (1.73):

$$\hat{D}^2(\hat{\bar{y}}_{\Omega_d}^{MGREG}) = \frac{1}{N_d^2} \hat{D}^2\left(\hat{y}_{\Omega_d}^{MGREG}\right). \tag{1.76}$$

This section also discusses the pseudo-empirical best linear unbiased predictor presented by Rao and Molina (2015). We assume a model belonging to area level models obtained by including the weights $\tilde{w}_{di} = \frac{w_{di}}{\sum_{j=1}^{n_d} w_{dj}}$ in the unit level model, thus (Rao and Molina, 2015, pp. 206–207):

$$\bar{y}_{dw} = \sum_{i=1}^{n_d} \tilde{w}_{di} y_{di} = \sum_{i=1}^{n_d} \tilde{w}_{di} \left(\mathbf{x}_{di}^T \boldsymbol{\beta} + v_d + e_{di} \right) = \bar{\mathbf{x}}_{dw} \boldsymbol{\beta} + v_d + \bar{e}_{dw}, \tag{1.77}$$

where $\mathbf{x}_{dw} = \sum_{i=1}^{n_d} \tilde{w}_{di} \mathbf{x}_{di}$, $\bar{e}_{dw} = \tilde{w}_{di} e_{dj}$ and w_{di} are the inverses of the first-order inclusion probabilities. In addition, $\bar{e}_{dw}$ has a distribution with the expectation value 0 and variance $\sigma_e^2 \delta_{dw}$ (where $\delta_{dw} = \sum_{i=1}^{n_d} \tilde{w}_{di}^2$). The best linear unbiased predictor $\mu_d = \bar{\mathbf{X}}_d^T \boldsymbol{\beta} + v_d$ under the post-aggregation model assumption is given by the formula (Rao and Molina, 2015, pp. 206–207):

$$\tilde{\mu}_{dw}^H = \bar{\mathbf{X}}_d^T \boldsymbol{\beta} + \gamma_{dw} \left(\bar{y}_{dw} - \bar{\mathbf{x}}_{dw} \boldsymbol{\beta} \right), \tag{1.78}$$

where $\gamma_{dw} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2 \delta_{dw}}$. Estimates of the parameters σ_v^2 and σ_e^2 for the type B model are obtained using the maximum likelihood method with constraints. In order to obtain a $\boldsymbol{\beta}$ estimate, it is necessary to obtain the best linear predictor of the random effect v_d:

$$\tilde{v}_{dw} \left(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2 \right) = \gamma_{dw} \left(\bar{y}_{dw} - \bar{\mathbf{x}}_{dw}^T \boldsymbol{\beta} \right), \tag{1.79}$$

and then the solution to the equation (Rao and Molina, 2015, p. 207) is:

$$\sum_{d=1}^{D} \sum_{i=1}^{n_d} w_{di} \mathbf{x}_{di} \left[y_{di} - \mathbf{x}_{di}^T \boldsymbol{\beta} - \tilde{v}_{dw} \left(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2 \right) \right].$$

We derive $\tilde{\boldsymbol{\beta}}_w$ from it, given by the formula:

$$\tilde{\boldsymbol{\beta}} \left(\sigma_v^2, \sigma_e^2 \right) = \left[\sum_{d=1}^{D} \sum_{i=1}^{n_d} \tilde{w}_{di} \mathbf{x}_{di} \left(\mathbf{x}_{di} - \gamma_{di} \bar{\mathbf{x}}_{dw} \right)^T \right]^{-1} \times \left[\sum_{d=1}^{D} \sum_{i=1}^{n_d} \tilde{w}_{di} \left(\mathbf{x}_{di} - \gamma_{di} \bar{\mathbf{x}}_{dw} \right) y_{di} \right]. \tag{1.80}$$

This estimator is ξ-unbiased. Replacing σ_e^2 and σ_v^2 by their estimator, we get the p-weighted estimator $\hat{\boldsymbol{\beta}}_w$. If we replace the $\boldsymbol{\beta}$ in (1.77) by the estimate $\hat{\boldsymbol{\beta}}_w$, we obtain a pseudo-empirical best linear unbiased predictor ($\hat{\mu}_{dw}^H$). Under the assumption of normality of the distribution of random effects and random components, the estimate $MSE\left(\hat{\mu}_{dw}^H \right)$ is given by the formula (You and Rao, 2002, p. 434):

$$M\hat{S}E \left(\hat{\mu}_{dw}^H \right) = g_{1dw} \left(\hat{\sigma}_v^2, \hat{\sigma}_e^2 \right) + g_{2dw} \left(\hat{\sigma}_v^2, \hat{\sigma}_e^2 \right) + 2 g_{3dw} \left(\hat{\sigma}_v^2, \hat{\sigma}_e^2 \right), \tag{1.81}$$

where

$$g_{1iw} \left(\sigma_v^2, \sigma_e^2 \right) = \gamma_{dw} \delta_{dw} \sigma_e^2, \tag{1.82}$$

$$g_{2iw} \left(\sigma_v^2, \sigma_e^2 \right) = \left(\bar{\mathbf{X}}_d - \gamma_{iw} \bar{\mathbf{x}}_{dw} \right)^T \Phi_w \left(\sigma_v^2, \sigma_e^2 \right) \left(\bar{\mathbf{X}}_d - \gamma_{dw} \bar{\mathbf{x}}_{dw} \right) \tag{1.83}$$

and

$$g_{3dw} \left(\sigma_v^2, \sigma_e^2 \right) = \gamma_{dw} (1 - \gamma_{dw})^2 \sigma_v^{-2} \sigma_e^{-2} h \left(\sigma_v^2, \sigma_e^2 \right). \tag{1.84}$$

In addition: $h\left(\sigma_v^2,\sigma_e^2\right)=\sigma_e^4\bar{V}_{vv}(\boldsymbol{\delta})+\sigma_v^4\bar{V}_{ee}(\boldsymbol{\delta})-2\sigma_e^2\sigma_v^2\bar{V}_{ve}(\boldsymbol{\delta})$, where $\bar{V}_{vv}$, $\bar{V}_{ee}$ and $\bar{V}_{ve}$ are the asymptotic variances and covariance of the estimators $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$. Also, $\Phi_w\left(\sigma_v^2,\sigma_e^2\right)$ is instead denoted by the variance-covariance matrix $\tilde{\boldsymbol{\beta}}\left(\sigma_v^2,\sigma_e^2\right)$ given by the formula:

$$\begin{aligned}\Phi_w\left(\sigma_v^2,\sigma_e^2\right) &= \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}\mathbf{x}_{di}\mathbf{z}_{di}^T\right)^{-1}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}\mathbf{z}_{di}\mathbf{z}_{di}^T\right)\left[\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}\mathbf{x}_{di}\mathbf{z}_{di}^T\right)^{-1}\right]^T\sigma_e^2\\ &+\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}\mathbf{x}_{di}\mathbf{z}_{di}^T\right)^{-1}\left[\sum_{d=1}^{D}\left(\sum_{i=1}^{n_d}\mathbf{z}_{di}\right)\left(\sum_{i=1}^{n_d}\mathbf{z}_{di}\right)^T\right]\left[\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}\mathbf{x}_{di}\mathbf{z}_{di}^T\right)^{-1}\right]^T\sigma_v^2, \quad (1.85)\end{aligned}$$

where $\mathbf{z}_{dj}=w_{dj}\left(\mathbf{x}_{dj}-\gamma_{dw}\bar{\mathbf{x}}_{dw}\right)$. Pseudo-EBLUP predictors were also considered by Prasad and Rao (1999), among others.

1.2. Superpopulation model and the steps of its construction

In this subsection, issues related to the key concept of the superpopulation model, which is central to the model-based approach, will be discussed in more detail. In the next subsections, a classification of superpopulation models will be presented with particular reference to the classes of models considered in this book, as well as the different steps in the process of building a superpopulation model.

1.2.1. Mixed models

In view of analyses conducted as part of this book, the classification of superpopulation models into models containing only fixed effects, random effects models and mixed models should be presented in more detail. Relevant to this classification of models are the concepts of fixed and random factor and fixed and random effect. We can speak of a random factor when a specific distribution is assumed for its levels, whereas we can speak of a fixed factor when the values are fixed (McCulloch, 2003, pp. 10–11). Following Biecek (2012), we can speak of the possibility to classify the effects of the levels of a variable as fixed effects when the number of levels of a trait is relatively small compared to the number of observations and does not change when the number of observations changes, and therefore when all variants of the trait are observed. In the case of fixed effects, the most common result of the analysis is a direct assessment of the effect value. Random effects, however, are treated as realisations of a random variable describing effects in the population. This is due to their abundance and, consequently, the possibility of there being some that are not observed. They cannot therefore be treated as model parameters. In the case of random effects, the aim is to assess the distribution of these effects, their variability in the population.

Fixed effects models were considered, among others, in the works of Borenstein et al. (2010) and Bramati and Croux (2007). Borenstein et al. (2010) use fixed effects models in meta-analyses. They emphasise that models of this class are applicable when two conditions are met. Firstly, we can assume that all studies are functionally identical, so that the conditions for conducting all studies are the same, e.g. the participants were recruited in the same way and the same people conduct the study. Secondly, the aim of the analysis is to determine a common effect size that would not be generalised beyond the (narrowly defined) population included in the analysis. Bramati and Croux (2007) used a fixed effects model in their analysis of private sector responses to fiscal policy. The authors used the national savings rate as the study variable, and the lagged national savings rate, the demand gap and the ratio of the population under 15 years and over 65 years, among others, as explanatory variables. Authors conducting analyses based on random effects models include, for example, Box and Tiao (1968), and Menegaki (2011). Box and Tiao (1968) consider the problem of estimating the mean in a random effect model from a Bayesian point of view. Menegaki (2011), nevertheless, used a random effects model in analyses of economic growth and renewable energy in 27 European countries. In her study, the author used panel data over a ten-year period (2007–2017).

Mixed models allow both fixed and random effects to be included in the model. They allow the analysis of data grouped by one or more classification variables. It should be noted that models belonging to this class can be applied to many types of data, including: cross-sectional and time series, repeated measures data and multivariate data. They also allow the modelling of multivariate data, and data with high variability and heterogeneity. The mixed model approach, according to Demidenko (2004), can be considered as a kind of compromise between the classical and Bayesian approaches.

Within the class of mixed models, we can make a division between linear and non-linear mixed models. All considerations and analyses presented in the remainder of this book apply to linear mixed models. In the case of non-linear mixed models, non-linearity may apply to all or only selected fixed and random effects alike. Models belonging to this class can, following Pinheiro and Bates (2000), be considered as extensions of non-linear regression models and linear mixed models. In the first case, by including random effects in the coefficients of the model, it is possible to account for between-group variation, correlation within groups, and in the second, the expected value of characteristics associated with random effects is a non-linear function. Non-linear mixed models have been considered in the work of Pinheiro and Bates (1995), Lindstrom and Bates (1990), Kuhn and Lavielle (2005), and Vonesh and Carter (1992).

The general non-linear mixed model has the following form (Lindstrom and Bates, 1990, pp. 674–675):

$$Y_{ij} = f(\boldsymbol{\phi_i}, \mathbf{x}_{ij}) + e_{ij}, \tag{1.86}$$

where Y_{ij} is the realisation of the random variable for the j-th response of the i-th unit, and f is a non-linear function of the vector of auxiliary variables denoted by $\mathbf{x}_{ij}$ and the parameters $\boldsymbol{\phi_i}$. In this model, we also assume a normal distribution of the random component e_{ij}. Furthermore, we can write the vector $\boldsymbol{\phi_i}$ as (Lindstrom and Bates, 1990, pp. 674–675):

$$\boldsymbol{\phi_i} = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{v}_i, \tag{1.87}$$

where $\boldsymbol{\beta}$ is the p-element vector of parameters (fixed effects) and $\mathbf{v}_i$ is the q-element vector of random effects. The $\mathbf{A}_i$ and $\mathbf{B}_i$ matrices of $n \times p$ and $n \times q$, respectively, are well-known matrices that simplify model specification. They allow different fixed-effects parameter values to be assigned for different groups of units or random effects to be prescribed to them or not. The above model for a single i-th response vector can be written as follows (Lindstrom and Bates, 1990, p. 675):

$$\mathbf{Y}_i = \boldsymbol{\eta}_i(\boldsymbol{\phi_i}) + \mathbf{e}_i, \tag{1.88}$$

where $\mathbf{e}_i \sim N\left(0, \sigma^2\boldsymbol{\Delta}\right)$ and the elements of the vector $\boldsymbol{\eta}_i(\boldsymbol{\phi_i})$ are the values of the function $f(\boldsymbol{\phi_i}, \mathbf{x}_{ij})$. In many cases, $\boldsymbol{\Delta}$ is a unitary matrix, however, with this matrix it is possible to take into account the covariance structure.

The general linear mixed model, the special cases of which are considered in this monograph, is given by the following formula (cf. Jiang, 2007, pp. 1–2):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \tag{1.89}$$

where $\mathbf{Y}$ – a random vector of values of the explanatory variable, $\mathbf{X}$, $\mathbf{Z}$ – known matrices of auxiliary variables, and $\boldsymbol{\beta}$ – a vector of unknown parameters. In this model, we assume that the random effects $\mathbf{v}$ and the random components $\mathbf{e}$ are independent, their expected values are equal to $\mathbf{0}$, and their variance-covariance matrices, respectively denoted as $\mathbf{G}(\boldsymbol{\delta})$ and $\mathbf{R}(\boldsymbol{\delta})$, depend on the variance component vector $\boldsymbol{\delta}$. The variance-covariance matrix $\mathbf{Y}$ is given by the formula (Littell et al., 2006, p. 736):

$$\mathbf{V}(\boldsymbol{\delta}) = \mathbf{Z}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^T + \mathbf{R}(\boldsymbol{\delta}). \tag{1.90}$$

The linear mixed model (1.89) can also be written in the following form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{v}_1 + \mathbf{Z}_2\mathbf{v}_2 + \cdots + \mathbf{Z}_h\mathbf{v}_h + \mathbf{e}, \tag{1.91}$$

where

$$D^2 \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_h \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} & \dots & \mathbf{G}_{1h} \\ \mathbf{G}_{21} & \mathbf{G}_{22} & \dots & \mathbf{G}_{2h} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_{h1} & \mathbf{G}_{h2} & \dots & \mathbf{G}_{hh} \end{bmatrix}.$$

It is important to note that in general, for $i \neq j$, it is possible that $\mathbf{G}_{ij} \neq \mathbf{0}$. Usually, it is additional assumed that the matrix $\mathbf{R}(\boldsymbol{\delta})$ has the form $\mathbf{R}(\boldsymbol{\delta}) = \sigma_e^2 \mathrm{diag}(a_i)$ for $1 \leqslant i \leqslant N$, where a_i is some known function of the auxiliary variables. It is therefore possible to account for the heteroscedasticity of the random components.

In small area estimation, model (1.91) and its special cases are considered, but when $v_1, v_2, \dots, v_h$ are independent, implying that $\mathbf{G}_{ij} = \mathbf{0}$ for each $i \neq j$ (cf. Rao and Molina, 2015, pp. 88–89). We propose to use the linear mixed model with correlated random effects vectors given by equation (1.91) in small area estimation without the additional assumption that $\mathbf{G}_{ij} = 0$ for each $i \neq j$.

Within the class of linear mixed models, Jiang (2007) distinguishes between Gaussian and non-Gaussian models. This classification is based on the assumption of normality of distribution. In the case of Gaussian models, the normality of the distribution of both effects and random components is assumed. In the case of non-Gaussian linear mixed models, it is assumed that the effects and random components are independent or uncorrelated and that their distributions are not normal. The joint distribution of the variable under study may not be fully specified. It should also be noted that the assumption of the normality of the distribution provides greater flexibility in modelling, whereas its absence provides resistance to failure to meet distribution assumptions.

Within linear mixed models, we can also distinguish between type A (area level models) and B (unit level models). It should be emphasised that unit level models require data at the level of individual units, while area level models use data at a higher aggregation level, e.g. for areas, groups, domains.

Amongst the models belonging to type A, we can distinguish the model considered by Fay and Herriot (1979) for single-period data:

$$\theta_d = \mathbf{x}_d^T \boldsymbol{\beta} + v_d, \tag{1.92}$$

$$\hat{\theta}_d = \theta_d + e_d, \tag{1.93}$$

where $d = 1, 2, \dots, D$, and $\hat{\theta}_d$ is some direct estimator of the characteristic θ_d. It should be added that in the classical Fay–Heriot model, the independence of the random components e_d and the random effects v_d is assumed. Furthermore, we assume that e_d have a distribution with an expected value of 0 and a known variance of σ_e^2. Similarly, for random effects v_d, a distribution with an expectation value of 0 and a variance of σ_v^2 is assumed. The vector of auxiliary variables for the domains is denoted by $\mathbf{x}_d$. It should be added that this model can allow accurate estimates to be obtained for small domains by using combined models for direct estimators, auxiliary variables and borrowing power from other domains. The model also allows data from different sources to be combined (Datta et al., 2005, p. 184; Rueda et al., 2010, p. 571). The Fay–Heriot model and its modifications have been used in estimating, among other things, per capita income, the unemployment rate in selected Canadian cities (Rao and Yu, 1994), the number of school-age children living in poverty (Lohr and Rao, 2009), and the value of the mean and kurtosis of household income (Jędrzejczak, 2011).

An example of a model belonging to the latter class is the model with a nested random component presented for single period data by Battese et al. (1988). This model for domain--specific random effects and longitudinal data has the following form (cf. Battese et al., 1988, pp. 29–30):

$$Y_{idt} = \mathbf{x}_{idt}^T \boldsymbol{\beta} + v_d + e_{idt}, \tag{1.94}$$

where $d = 1, 2, \ldots, D$, $i = 1, 2, \ldots, N$, and $t = 1, 2, \ldots, M$. As in the model given by the formulas (1.92) and (1.93), we make the assumptions that e_{idt} and v_d have distributions with expectation values of 0 and variances of σ_e^2 and σ_v^2, respectively, and therefore $v_d \sim iid\left(0, \sigma_v^2\right)$ and $e_{idt} \sim iid\left(0, \sigma_e^2\right)$. In addition, the variance-covariance matrices for the effects and random components have the following forms:

$$\mathbf{G}(\boldsymbol{\delta}) = \sigma_{v_d}^2 \mathbf{I}_{D\times D}, \tag{1.95}$$

$$\mathbf{R}(\boldsymbol{\delta}) = \sigma_e^2 \mathrm{diag}(a_{it}) \tag{1.96}$$

for $1 \leqslant i \leqslant N$ and $1 \leqslant t \leqslant M$, where a_{it} is some known function of the auxiliary variables. Thus, for the model (1.94), we also have:

$$E\left(Y_{idt}\right) = \mathbf{x}_{idt}^T \boldsymbol{\beta} \tag{1.97}$$

and

$$Cov_\xi\left(Y_{idt}, Y_{i'd't'}\right) = \begin{cases} \sigma_e^2 + \sigma_v^2 & \text{when } i = i' \wedge d = d' \wedge t = t', \\ \sigma_v^2 & \text{for } i \neq i' \wedge d = d', \\ 0 & \text{for } d \neq d'. \end{cases} \tag{1.98}$$

For the above model, the $\mathbf{Z}$ matrix can be written as:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1}_{N_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{N_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{N_D} \end{bmatrix}_{NM \times D},$$

where $\mathbf{1}_{N_d}$ is a vector of 1's of dimension $N_dM \times 1$ while the matrix $\mathbf{V}(\boldsymbol{\delta})$ is of the form:

$$\mathbf{V}(\boldsymbol{\delta}) = \underset{1 \leqslant d \leqslant D}{\operatorname{diag}} \mathbf{V}_d = \underset{1 \leqslant d \leqslant D}{\operatorname{diag}} \left(\sigma_{v_d}^2 \mathbf{1}_{N_dM} \mathbf{1}_{N_dM}^T + \sigma_e^2 \mathbf{I}_{N_dM \times N_dM}\right). \tag{1.99}$$

The model proposed by Battese et al. (1988) or some modifications of it have found application, among others, in forecasting crop areas using geodetic and satellite data (Battese et al., 1988). The issue of prediction using the above model was also considered in the work of Prasad and Rao (1990), where the problem of estimating the mean squared error was also considered, Torabi and Rao (2010), where the use of pseudo-EBLUP was considered, and Rivest et al. (2016), where combining functions were used.

Among the single random effect models, besides the model with a nested random component presented above, we can distinguish a model with a random slope. This model with a domain-specific random effect for longitudinal data has the form (for data from a single period, cf. Dempster et al., 1981, pp. 342–344):

$$Y_{idt} = (\beta_1 + v_d)x_{idt} + \beta_0 + e_{idt}, \tag{1.100}$$

where $i = 1,2,\ldots,N$, $d = 1,2,\ldots,D$, $t = 1,2,\ldots,M$, $v_d \sim iid\left(0,\sigma_d^2\right)$, and $e_{idt} \sim iid\left(0,\sigma_e^2\right)$. The matrices $\mathbf{G}(\boldsymbol{\delta})$ and $\mathbf{R}(\boldsymbol{\delta})$ have the form given by the formulas (1.95) and (1.96), and therefore as in the case with a nested random component, the matrix of auxiliary variables $\mathbf{Z}$ can be written as:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_D \end{bmatrix}_{NM \times D},$$

where $i = 1,2,\ldots,N$, $d = 1,2,\ldots,D$, $t = 1,2,\ldots,M$, and $\mathbf{x}_d$ has dimension $N_dM \times 1$. The variance-covariance matrix of the vector of random variables under consideration is given in this case by the following formula:

$$\mathbf{V}(\boldsymbol{\delta}) = \underset{1 \leqslant d \leqslant D}{\operatorname{diag}} \mathbf{V}_d = \underset{1 \leqslant d \leqslant D}{\operatorname{diag}} \left(\sigma_{v_d}^2 \mathbf{x}_d \mathbf{x}_d^T + \sigma_e^2 \mathbf{I}_{N_dM \times N_dM}\right). \tag{1.101}$$

Similarly, both models can be written, for example, for time- or profile-specific random effects, thus v_t or v_i. This model, with modifications, has found applications in, for example, estimating students' average grade points after the first year of law (Dempster et al., 1981) and glucose tolerance (Reinsel, 1984).

Linear mixed models with two or more random effects also allow us to distinguish classes of models that take into account the presence or lack of a correlation between random effects. These models will be discussed in more detail in the next two subsections 1.2.2 and 1.2.3.

1.2.2. Special cases of linear mixed models with uncorrelated random effects

This subsection of the monograph will present selected special cases of linear mixed models with two, three and four random effects. For models with two random effects, how Biecek (2012) points out, we can consider cases where the random effects involve the same or different grouping variables. For classes of models with more than two effects, we can also consider several grouping variables.

The first model presented is the model with two uncorrelated random effects specific to the two grouping variables. Following Stukel and Rao (1999) this model with domain-specific and profile-specific random effects is given by the formula:

$$Y_{idt} = \sum_{k=1}^{p} \beta_k x_{k,idt} + v_d + v_{id} + e_{idt}, \tag{1.102}$$

where $i = 1,2,\ldots,N$, $d = 1,2,\ldots,D$, $t = 1,2,\ldots,M$, $v_d \sim iid\big(0,\sigma^2_{v_d}\big)$, $v_{id} \sim iid\Big(0,\sigma^2_{v_id}\Big)$, and $e_{idt} \sim iid\big(0,\sigma^2_e\big)$. Furthermore, for the above model, the following occurs:

$$Cov_{\xi}\left(Y_{id},Y_{i'd'}\right) = \begin{cases} \sigma^2_e + \sigma^2_{v_d} + \sigma^2_{v_{id}} & \text{when } i = i' \wedge d = d' \wedge t = t', \\ \sigma^2_{v_d} + \sigma^2_{v_{id}} & \text{when } i = i' \wedge d = d' \wedge t \neq t', \\ \sigma^2_{v_d} & \text{when } i \neq i' \wedge d = d', \\ 0 & \text{in other cases.} \end{cases} \tag{1.103}$$

We can decompose the matrix of auxiliary variables $\mathbf{Z}$ for the above model into two component matrices $\mathbf{Z}^{(D)}$ and $\mathbf{Z}^{(ID)}$:

$$\mathbf{Z} = \Big[\mathbf{Z}^{(D)} \quad \mathbf{Z}^{(ID)}\Big]_{NM\times(N+D)}, \tag{1.104}$$

where the component matrices are of the following form:

$$\mathbf{Z}^{(D)} = \begin{bmatrix} \mathbf{1}_{MN_1} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{1}_{MN_d} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{1}_{MN_D} \end{bmatrix}_{NM\times D} \tag{1.105}$$

and

$$\mathbf{Z}^{(ID})= \begin{bmatrix} \mathbf{1}_M & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{1}_M & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{1}_M \end{bmatrix}_{NM\times N} . \tag{1.106}$$

The matrix $\mathbf{V}(\boldsymbol{\delta})$ in this case can be written as $\mathbf{V}(\boldsymbol{\delta}) = \operatorname*{diag}_{1\leqslant d\leqslant D} \mathbf{V}_d$, where the matrix $\mathbf{V}_d$ is of the form:

$$\mathbf{V}_d = \begin{bmatrix} \mathbf{V}_{11} & \cdots & \mathbf{V}_{1i} & \cdots & \mathbf{V}_{1N_d} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{V}_{i1} & \cdots & \mathbf{V}_{ii} & \cdots & \mathbf{V}_{1N_d} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{V}_{N_d1} & \cdots & \mathbf{V}_{N_di} & \cdots & \mathbf{V}_{N_dN_d} \end{bmatrix}_{MN_d\times MN_d} . \tag{1.107}$$

The variance-covariance matrix of the random variables for i-th element of the population at different periods $\mathbf{V}_{ii}$ is given by the formula:

$$\mathbf{V}_{ii} = \begin{bmatrix} \sigma_e^2+\sigma_{v_d}^2+\sigma_{v_{id}}^2 & \cdots & \sigma_{v_d}^2+\sigma_{v_{id}}^2 & \cdots & \sigma_{v_d}^2+\sigma_{v_{id}}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sigma_{v_d}^2+\sigma_{v_{id}}^2 & \cdots & \sigma_e^2+\sigma_{v_d}^2+\sigma_{v_{id}}^2 & \cdots & \sigma_{v_d}^2+\sigma_{id}^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sigma_{v_d}^2+\sigma_{v_{id}}^2 & \cdots & \sigma_{v_d}^2+\sigma_{v_{id}}^2 & \cdots & \sigma_e^2+\sigma_{v_d}^2+\sigma_{v_{id}}^2 \end{bmatrix}_{M\times M} , \tag{1.108}$$

and the variance-covariance matrix between the random variables for the i-th and i'-th elements of the population at different periods has the form:

$$\mathbf{V}_{ii'} = \sigma_{id}^2\mathbf{1}_M\mathbf{1}_M^T. \tag{1.109}$$

Another of the cases considered is a model with two uncorrelated random effects specific to one grouping variable – the domain. It is given by the following formula (Krzciuk, 2020, p. 20):

$$Y_{idt} = (\beta_1+v_{2d})x_{idt}+\beta_0+v_{1d}+e_{idt}, \tag{1.110}$$

where $i=1,2,\ldots,N$, $d=1,2,\ldots,D$, $t=1,2,\ldots,M$, $v_{1d}\sim iid\left(0,\sigma_{v_{1d}}^2\right)$, $v_{2d}\sim iid\left(0,\sigma_{v_{2d}}^2\right)$, and

$e_{idt} \sim iid\left(0, \sigma_e^2\right)$. For this model, the $\mathbf{G}(\boldsymbol{\delta})$ matrix is a block-diagonal matrix:

$$\mathbf{G}(\boldsymbol{\delta}) = D^2 \begin{bmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{1d} \\ v_{2d} \\ \vdots \\ v_{1D} \\ v_{2D} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_1(\boldsymbol{\delta}) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_2(\boldsymbol{\delta}) & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \dots & \dots & \mathbf{G}_D(\boldsymbol{\delta}) \end{bmatrix}_{2D \times 2D}, \tag{1.111}$$

where the submatrix of the $\mathbf{G}$ matrix for the domain can be written as:

$$\mathbf{G}_d(\boldsymbol{\delta}) = \begin{bmatrix} \sigma_{v_{1d}}^2 & 0 \\ 0 & \sigma_{v_{2d}}^2 \end{bmatrix}. \tag{1.112}$$

The variance-covariance matrix $\mathbf{R}(\boldsymbol{\delta})$, in this and the next of the special cases presented, has the form as in the previous models, i.e. given by the formula (1.96). When two random effects are included in the model, the matrix $\mathbf{Z}$ has a slightly more complex form:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1}_{N_1} & \mathbf{x}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{N_2} & \mathbf{x}_2 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{N_D} & \mathbf{x}_D \end{bmatrix}_{NM \times 2D}, \tag{1.113}$$

where $\mathbf{1}_{N_d}$ and $\mathbf{x}_d$ have dimension $N_d M \times 1$. The matrix $\mathbf{V}$ in this case can be written as the sum of the components of $\mathbf{ZGZ}^{-1}$ for models with a nested random component and a random slope and the matrix $\mathbf{R}$, thus as:

$$\mathbf{V}(\boldsymbol{\delta}) = \underset{1 \leqslant d \leqslant D}{\operatorname{diag}} \mathbf{V}_d = \underset{1 \leqslant d \leqslant D}{\operatorname{diag}} \left(\sigma_{v_{1d}}^2 \mathbf{1}_{N_d M} \mathbf{1}_{N_d M}^T + \sigma_{v_{2d}}^2 \mathbf{x}_d \mathbf{x}_d^T + \sigma_e^2 \mathbf{I}_{N_d M \times N_d M}\right). \tag{1.114}$$

We can also consider more complex models containing three or more random effects. The first model considered containing three random effects can be written as:

$$Y_{idt} = (\beta_1 + v_{2d}) x_{idt} + \beta_0 + v_{1d} + v_t + e_{idt}, \tag{1.115}$$

where $v_{1d} \sim iid\left(0, \sigma_{v_{1d}}^2\right)$, $v_{2d} \sim iid\left(0, \sigma_{v_{2d}}^2\right)$, and $e_{idt} \sim iid\left(0, \sigma_e^2\right)$. In addition to domain-specific effects, a random effect for time v_t is included in this model. The variance-covariance matrix of the random effects for the above model has the following form:

$$\mathbf{G}(\boldsymbol{\delta}) = \begin{bmatrix} \mathbf{G}_1(\boldsymbol{\delta}) & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_2(\boldsymbol{\delta}) & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_D(\boldsymbol{\delta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G}_T(\boldsymbol{\delta}) \end{bmatrix}_{(2D+M)\times(2D+M)}, \tag{1.116}$$

where the domain-specific random effects submatrix is given by the formula (1.112), and for time we can write as:

$$\mathbf{G}_T(\boldsymbol{\delta}) = \sigma^2_{v_t}\mathbf{I}_{M\times M}. \tag{1.117}$$

We can decompose the matrix of auxiliary variables $\mathbf{Z}$ for the above model into two component matrices $\mathbf{Z}^{(D)}$ and $\mathbf{Z}^{(T)}$:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}^{(D)} & \mathbf{Z}^{(T)} \end{bmatrix}_{NM\times(2D+M)}, \tag{1.118}$$

the matrix associated with the domain-specific random effects $\mathbf{Z}^{(D)}$ given by the formula (1.113) and the matrix $\mathbf{Z}^{(T)}$, corresponding to the third random effect:

$$\mathbf{Z}^{(T)} = \begin{bmatrix} \mathbf{Z}^{(T)}_1 \\ \mathbf{Z}^{(T)}_2 \\ \vdots \\ \mathbf{Z}^{(T)}_D \end{bmatrix}_{NM\times M}, \tag{1.119}$$

where

$$\mathbf{Z}^{(T)}_d = \begin{bmatrix} \mathbf{1}_{d1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{d2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{dM} \end{bmatrix}_{N_dM\times M}, \tag{1.120}$$

and $\mathbf{1}_{dt}$ has the dimension $N_d \times 1$.

The second variant of this model considered is given by the following equation:

$$Y_{idt} = (\beta_1 + v_{2d} + v_t)x_{idt} + \beta_0 + v_{1d} + e_{idt}, \tag{1.121}$$

where $v_{1d} \sim iid\left(0, \sigma^2_{v_{1d}}\right)$, $v_{2d} \sim iid\left(0, \sigma^2_{v_{2d}}\right)$, and $e_{idt} \sim iid\left(0, \sigma^2_e\right)$. Again, it is possible to decompose the matrix of auxiliary variables $\mathbf{Z}$ given by the formula (1.118) and the matrix $\mathbf{Z}^{(T)}$ has the form (1.119), where the submatrices can be written as:

$$\mathbf{Z}^{(T)}_d = \begin{bmatrix} \mathbf{x}^{(T)}_{d1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}^{(T)}_{d2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}^{(T)}_{dM} \end{bmatrix}_{N_dM\times M}, \tag{1.122}$$

where $\mathbf{x}_{dt}^{(T)}$ has the dimension $N_d \times 1$. Then the matrix $\mathbf{G}(\boldsymbol{\delta})$ is given by the formula (1.116).

The last of the mixed models with uncorrelated random effects presented in this subsection is the model with four random effects:

$$Y_{idt} = (\beta_1 + v_{2d} + v_{2t})x_{idt} + \beta_0 + v_{1d} + v_{1t} + e_{idt}, \tag{1.123}$$

where $v_{1d} \sim iid\left(0, \sigma^2_{v_{1d}}\right)$, $v_{2d} \sim iid\left(0, \sigma^2_{v_{2d}}\right)$, and $e_{idt} \sim iid\left(0, \sigma^2_e\right)$. The model (1.121) considered in this paragraph is extended with a second time-specific effect. The variance-covariance matrix of the random effects for this model is given by the formula:

$$\mathbf{G}(\boldsymbol{\delta}) = \begin{bmatrix} \mathbf{G}_1(\boldsymbol{\delta}) & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_2(\boldsymbol{\delta}) & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_D(\boldsymbol{\delta}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G}_T(\boldsymbol{\delta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{G}_T(\boldsymbol{\delta}) \end{bmatrix}_{(2D+2M)\times(2D+2M)}. \tag{1.124}$$

The submatrix corresponding to domain-specific effects is of the same form as for the other models, assuming a lack of correlation, and therefore follows the formula (1.112). The matrix $\mathbf{G}_T(\boldsymbol{\delta})$ has a form analogous to the submatrix for domain-specific effects:

$$\mathbf{G}_T(\boldsymbol{\delta}) = \begin{bmatrix} \sigma^2_{v_{1t}} & 0 \\ 0 & \sigma^2_{v_{2t}} \end{bmatrix}. \tag{1.125}$$

For this model, it is also possible to decompose the $\mathbf{Z}$ matrix according to the formula (1.118), however, it has dimensions $NM \times (2D + 2M)$. We can write the submatrix for time as (1.119), where $\mathbf{Z}_d^{(T)}$ has form:

$$\mathbf{Z}_d^{(T)} = \begin{bmatrix} \mathbf{1}_{d1}^{(T)} & \mathbf{x}_{d1}^{(T)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{d2}^{(T)} & \mathbf{x}_{d2}^{(T)} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{dM}^{(T)} & \mathbf{x}_{dM}^{(T)} \end{bmatrix}_{N_dM\times 2M}. \tag{1.126}$$

1.2.3. Special cases of linear mixed models with correlated random effects

When discussing models belonging to the class of linear mixed models, we can also consider the occurrence of correlations between random effects and between vectors of random effects in the case of models including two or more random effects.

When considering models with correlated random effects, attention should be paid to models in which the spatial correlation of random effects is assumed. In such a case, the vector of correlated random effects for the domains $\mathbf{v}$, assuming a simultaneous spatial autoregressive process (SAR process), is given by the following formula (Cressie, 1993):

$$\mathbf{v} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u}, \tag{1.127}$$

where $\mathbf{u}$ is a D-element vector of independent random effects with variance σ_u^2, and ρ is an unknown parameter. The spatial weights matrix $\mathbf{W}$ is of dimension $D \times D$, since correlation is assumed between domains rather than between population elements. It should be noted that the proximity of domains can be considered not only in a geographic sense, but also in an economic sense, using variables such as the unemployment rate or the value of investments (Pietrzak, 2010). It should be added that the rows of the $\mathbf{W}$ matrix are usually standardised. The problem of defining weight matrices has been widely presented in a book edited by Suchecki (2010).

The variance-covariance matrix of the random effects $\mathbf{G}$ is given in this case by the formula (Molina et al., 2008, p. 444):

$$\mathbf{G} = \sigma_u^2\left[(\mathbf{I} - \rho\mathbf{W})\left(\mathbf{I} - \rho\mathbf{W}^T\right)\right]^{-1}. \tag{1.128}$$

The $\mathbf{R}$ matrix can be written as:

$$\mathbf{R} = \operatorname{diag}\left(\sigma_e^2\right). \tag{1.129}$$

By substituting (1.128) and (1.129) into (1.90), we obtain a matrix $\mathbf{V}$ which has the form (Pratesi and Salvati, 2008, p. 116):

$$\mathbf{V} = \mathbf{ZGZ}^T + \mathbf{R} = \mathbf{Z}\sigma_u^2\left[(\mathbf{I} - \rho\mathbf{W})\left(\mathbf{I} - \rho\mathbf{W}^T\right)\right]^{-1}\mathbf{Z}^T + \sigma_e^2\mathbf{I}. \tag{1.130}$$

The problem of spatial correlation of random effects was considered, among others, in the studies of Pratesi and Salvati (2008) and Petrucci and Savalati (2004), for type A and B models.

The first model proposed that takes into account the correlation of random effects vectors, which is a special case of model (1.93), is a model containing two correlated domain-specific random effects (Krzciuk, 2020, p. 20):

$$Y_{idt} = (\beta_1 + v_{2d}^*)x_{idt} + \beta_0 + v_{1d}^* + e_{idt}, \tag{1.131}$$

where v_{1d}^* and v_{2d}^* are domain-specific random effects, $v_{1d}^* \sim iid\left(0, \sigma_{v_{1d}^*}^2\right)$ for $d = 1, 2, \ldots, D$, $v_{2d}^* \sim iid\left(0, \sigma_{v_{2d}^*}^2\right)$ for $d = 1, 2, \ldots, D$, $cor\left(v_{1d}^*, v_{2d}^*\right) = \rho$ for $d = 1, 2, \ldots, D$, and $e_{idt} \sim iid\left(0, \sigma_e^2\right)$. The $\mathbf{G}(\boldsymbol{\delta})$ matrix is also in this case a block-diagonal matrix, as in the model with uncorrelated random effects (1.111), however, the submatrix for the domain is given by the formula:

$$\mathbf{G}_d^* = \begin{bmatrix} \sigma_{v_{1d}^*}^2 & \rho\sigma_{v_{1d}^*}\sigma_{v_{2d}^*} \\ \rho\sigma_{v_{1d}^*}\sigma_{v_{2d}^*} & \sigma_{v_{2d}^*}^2 \end{bmatrix}. \tag{1.132}$$

The variance-covariance matrix $\mathbf{Y}$ can be written as:

$$\begin{aligned}\mathbf{V}^*(\boldsymbol{\delta}) &= \operatorname*{diag}_{1\leqslant d\leqslant D} \mathbf{V}_d = \\ &= \operatorname*{diag}_{1\leqslant d\leqslant D} \left(\sigma^2_{v^*_{1d}} \mathbf{1}_{N_dM}\mathbf{1}^T_{N_dM} + \sigma^2_{v^*_{2d}} \mathbf{x}_d\mathbf{x}_d^T + \rho\sigma_{v^*_1}\sigma_{v^*_2}\left(\mathbf{1}_{N_dM}\mathbf{x}_d^T + \mathbf{x}_d\mathbf{1}^T_{N_dM}\right) + \sigma_e^2\mathbf{I}_{N_dM\times N_dM}\right),\end{aligned} \tag{1.133}$$

where $\mathbf{x}_d$ is a vector of auxiliary variable values of dimension $N_dM \times 1$. Compared to the $\mathbf{V}(\boldsymbol{\delta})$ matrix for the (1.110) model, it has been supplemented with one additional component containing the ρ parameter and taking into account the correlation between random effects.

We can also consider more complex models containing three or more random effects. In these cases, it is also possible to take into account the correlation between the random effects. We consider two variants of the model containing three random effects. The first of these proposals can be written as:

$$Y_{idt} = (\beta_1 + v^*_{2d})x_{idt} + \beta_0 + v^*_{1d} + v_t + e_{idt}, \tag{1.134}$$

where $v_t \sim iid\left(0,\sigma^2_{v_t}\right)$, $v^*_{1d} \sim iid\left(0,\sigma^2_{v^*_{1d}}\right)$, $v^*_{2d} \sim iid\left(0,\sigma^2_{v^*_{2d}}\right)$, $cor\left(v^*_{1d},v^*_{2d}\right) = \rho$, and $e_{idt} \sim iid\left(0,\sigma^2_e\right)$ for $d = 1,2,\ldots,D$. The third random effect included in the model is the random effect for time v_t. The variance-covariance matrix of the random effects for this model has the following form:

$$\mathbf{G}^{**}(\boldsymbol{\delta}) = \begin{bmatrix} \mathbf{G}^*_1(\boldsymbol{\delta}) & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^*_2(\boldsymbol{\delta}) & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}^*_D(\boldsymbol{\delta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G}_T(\boldsymbol{\delta}) \end{bmatrix}_{(2D+M)\times(2D+M)}, \tag{1.135}$$

where the domain-specific random effects submatrix is given by the formula (1.132), and for time, we can write it as:

$$\mathbf{G}_T(\boldsymbol{\delta}) = \sigma^2_{v_t}\mathbf{I}_{M\times M}. \tag{1.136}$$

The $\mathbf{Z}$ auxiliary variable matrix for the above model can be decomposed into two component matrices. It thus has the form given by the formula (1.118), where the matrix associated with the domain-specific random effects and the matrix $\mathbf{Z}^{(T)}$, which corresponds to the third random effect, can be written with the formulas (1.113) and (1.119), respectively.

The second variant of this model considered is given by the following equation:

$$Y_{idt} = (\beta_1 + v^*_{2d} + v_t)x_{idt} + \beta_0 + v^*_{1d} + e_{idt}, \tag{1.137}$$

where we make assumptions about the distributions of the random effects and the random component as in the (1.134) model. Once again, it is possible to decompose the matrix of the auxiliary variables $\mathbf{Z}$ and the matrix $\mathbf{Z}^{(T)}$ has the form (1.119), where the submatrices are given by the formula (1.122). The matrix $\mathbf{G}(\boldsymbol{\delta})$ is given by the formula (1.135).

The last of the special cases of the mixed model with correlated random effects vectors presented in this subsection is the model with four random effects:

$$Y_{idt} = (\beta_1 + v^*_{2d} + v^*_{2t})x_{idt} + \beta_0 + v^*_{1d} + v^*_{1t} + e_{idt}, \tag{1.138}$$

where v^*_{1d}, v^*_{2d}, v^*_{1t}, and v^*_{2t} are random effects specific to the domain and time, respectively. In addition $v^*_{1d} \sim iid\left(0, \sigma^2_{v^*_{1d}}\right)$ for $d = 1,2,\ldots,D$, $v^*_{2d} \sim iid\left(0, \sigma^2_{v^*_{2d}}\right)$ for $d = 1,2,\ldots,D$, $v^*_{1t} \sim iid\left(0, \sigma^2_{v^*_{1t}}\right)$ for $t = 1,2,\ldots,M$, $v^*_{2t} \sim iid\left(0, \sigma^2_{v^*_{2t}}\right)$ for $t = 1,2,\ldots,M$, $cor\left(v^*_{1d}, v^*_{2d}\right) = \rho$ for $d = 1,2,\ldots,D$, $cor\left(v^*_{1t}, v^*_{2t}\right) = \rho'$ for $t = 1,2,\ldots,M$, and $e_{idt} \sim iid\left(0, \sigma^2_e\right)$. The models considered in this paragraph have been supplemented with a second time-specific effect. In addition, the correlation between both domain-specific and time-specific effects has been taken into account. The variance-covariance matrix of the random effects for this case is given by the formula (1.121). The submatrix corresponding to the domain-specific effects is of the same form as for the other models assuming correlation, and therefore follows the formula (1.132). The matrix $\mathbf{G}_T$ is supplemented with off-major diagonal elements, taking correlation into account, and thus has a form analogous to the submatrix of the matrix $\mathbf{G}^*_d$ given by the formula (1.132):

$$\mathbf{G}_T = \begin{bmatrix} \sigma^2_{v^*_{1t}} & \rho' \sigma_{v^*_{1t}} \sigma_{v^*_{2t}} \\ \rho' \sigma_{v^*_{1t}} \sigma_{v^*_{2t}} & \sigma^2_{v^*_{2t}} \end{bmatrix}. \tag{1.139}$$

For this model, it is also possible to decompose the $\mathbf{Z}$ matrix according to the formula (1.118) but it has dimensions $N \times (2D + 2M)$. We can write the submatrix for time as (1.119), where the component matrices are given by the formula (1.126). The matrix corresponding to domain--specific effects remains consistent with the formula (1.132). Although some special cases of mixed models with correlated random effects vectors have been considered in the work of Dumont et al. (2014), Menec et al. (2004), and Ogungbenro et al. (2008), to the best of our knowledge, the class of linear mixed models with correlated random effects (1.93) is a new proposal, and its special cases have not been used in small area estimation to date.

1.2.4. Steps of model construction

One of the most important elements in the modelling approach is the process of model building. This subsection will discuss the different steps in this process – model specification, estimation and verification.

The first phase of model building is model specification. This phase consists primarily of defining the purpose of building the model and determining the dependent and auxiliary variables.

According to Pawłowski (1969), when selecting variables, indications of a theory concerning the phenomenon under study may be helpful. When theory does not provide a sufficient basis for determining the characteristics that should be included in the model, the author points to basing the decision on empirical material, other studies and computational experiments. It is also emphasised that the variables used in the model should be clearly defined and have an economic interpretation (Sobczyk, 2012, p. 12). Numerous statistical and econometric methods can be used in the selection of variables for the model. Among the classical methods, we can distinguish: the Hellwig (1969) method (optimal choice of predictors), the method proposed by Pawłowski (1981) (elimination of quasi-constant variables), and the graph method proposed by Bartosiewicz (1973) (sequential methods of variable selection).

At this stage, the sources of data to be used in the model are also identified. These data, depending on the problem under investigation, may take the form of cross-sectional data, time series, or longitudinal data. It should be noted that the statistical material, i.e. the set of data obtained by observation – one of the first steps of statistical investigation – can be divided into two categories: primary and secondary. Primary data are data collected directly for the implementation of a specific survey or statistical programme, e.g. data collected during the National Census. Secondary data, however, can be referred to when data were collected for other purposes, but were also used in a statistical survey (cf. Sobczyk, 2004, pp. 20–21). An example of secondary data is data from official registers.

The next element in the specification of the model is the choice of its analytical form. In the literature, we can find this step as a separate step in the model building process (cf. Strahl et al., 2004, pp. 29–30). This choice is often made on the basis of the results of other studies, computational experiments conducted, analysis of graphs of the study variable and explanatory variables, or properties of mathematical functions (cf. Guzik, 2008, p. 25; Sobczyk, 2012, p. 12). Non-statistical information on the phenomenon under study and the regularities that are associated with it can also be of great importance in determining the analytical form of the model.

Verbecke and Molenberghs (2000), in the case of linear mixed models, thus the class of models considered in this book, distinguish two steps of model specification. These are the choice of the mean value structure $(\mathbf{X}\boldsymbol{\beta})$ and the covariance structure. It should be added that the first of these stages is related to the selection of the auxiliary variables and thus the determination of the fixed effects in the model, while the second is related to the specification of the random effects to be included in the model. It is therefore related to the selection of grouping variables. The authors point out that these stages are not independent of each other unless robust methods are used. Adequate identification of the covariance structure allows correct inference of parameters for fixed effects.

When using computational experiments, therefore one of the heuristic methods, it is important to select a criterion on the basis of which the goodness of fit of the selected model form to the considered data set will be determined. Among these, we can distinguish the information criteria of Akaike (1973) (AIC), Schwartz (1978) (BIC), Hannan and Quinn (1979) (HQIC), and Bozdogan (1987) (CAIC), which are modifications of the AIC criterion. They are special cases of the generalised information criterion (GIC) having the form (Biecek, 2012, p. 123):

$$GIC = -2\ln L(M) + h|M|, \tag{1.140}$$

where L is the reliability function for the model M and $|M|$ is the number of model parameters. For the most commonly used AIC and BIC criteria, the parameter h takes the value of 2 and $\ln(n)$, respectively, where n denotes the number of observations. When comparing models based on the GIC criterion and its special cases, we choose the model with the smallest value of the measure (Biecek, 2012, p. 123). It should be added that these criteria can also be applied to the mixed models considered in this monograph.

The next stage of model construction is the estimation of its parameters. The decision on the method of parameter estimation results primarily from the choice of the analytical form of the model made at the stage of its specification. This book will discuss selected methods of estimating the parameters of the model, relevant in the context of the issues addressed in this monograph.

Methods for estimating model parameters, important in the context of the class of models analysed in this monograph, are the maximum likelihood method (ML) and the restricted maximum likelihood method (REML). The maximum likelihood method, following Jiang (2007), was first used in the 1920s by Fisher (1922). However, the method was only used to estimate the parameters of mixed models in the late 1960s by Hartley and Rao (1967). Under the assumption that the test variable $\mathbf{Y}$ has a multivariate normal distribution ($\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$), the likelihood function is given by the formula:

$$L = -\frac{1}{\sqrt{(2\pi)^n|\mathbf{V}_{ss}|}} \exp\left(-\frac{1}{2}\left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right)\mathbf{V}_{ss}^{-1}\left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right)\right), \tag{1.141}$$

and its logarithm has the form:

$$l = -\frac{1}{2}\left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right)\mathbf{V}_{ss}^{-1}\left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right) - \frac{1}{2}n\ln(2\pi) - \frac{1}{2}\ln\left(|\mathbf{V}_{ss}|\right). \tag{1.142}$$

In order to obtain an assessment of the vector $\boldsymbol{\delta}$, thus the vector of unknown parameters on which the variance-covariance matrix depends, by means of the scoring algorithm, it is necessary to determine the vector of first derivatives of the logarithm of the l function and the matrix of expected values of the $-l$ function's hessian. This procedure is outlined in more detail by Rao and Molina (2015). This algorithm yields the maximum-likelihood estimators $\hat{\boldsymbol{\delta}}_{ML}$ and $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\delta}}_{ML}\right)$. The asymptotic variance-covariance matrices of these estimators are given by the following formulae:

$$\breve{D}^2_{\xi}\left(\hat{\boldsymbol{\beta}}_{ML}\right) = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}, \tag{1.143}$$

$$\breve{\boldsymbol{\delta}}^2_{\xi}\left(\hat{\boldsymbol{\delta}}_{ML}\right) = \mathbf{I}^{-1}(\boldsymbol{\delta}). \tag{1.144}$$

The Newthon–Raphson algorithm can also be used to determine the assessment of $\boldsymbol{\delta}$. In this algorithm, however, negative values of $\mathbf{s}$ are possible, and it does not provide, as Biecek (2012) points out, a guarantee of convergence to a local maximum. It should also be noted that the estimator of the generalised least-squares method is used to assessment $\boldsymbol{\beta}$ in this method, which is also a drawback of this approach. When discussing the above method, it can also be added that the equivalent of maximising the likelihood function, or its logarithm, is to minimise the deviance given by the following formula (Biecek, 2012, p. 150):

$$-2l = -2(\ln L).$$

The origins of the restricted maximum-likelihood method date back to the early 1960s and the work of Thompson (1962). In this method, we consider a transformation of the vector $\mathbf{Y}$ such that:

$$\mathbf{Y}^*_s = \mathbf{A}^T\mathbf{Y}_s, \tag{1.145}$$

where $\mathbf{A}$ is an arbitrary matrix of order $n-p$, with n rows and $n-p$ columns, orthogonal to the matrix of auxiliary variables $\mathbf{X}_s$ $\left(\mathbf{A}^T\mathbf{X}_s = \mathbf{0}\right)$. It should be added that when $\mathbf{Y}_s$ has a multivariate normal distribution, after transformation $\mathbf{Y}^*_s$ has an $(n-p)$-dimensional normal distribution, with vector of expected values $\mathbf{0}$ and variance-covariance matrix $\mathbf{A}^T\mathbf{V}_{ss}\mathbf{A}$ (Rao and Molina, 2015, pp. 102–103). The logarithm of the restricted likelihood function is given in this case by the formula:

$$\ln L_R = -\frac{1}{2}\mathbf{Y}^T_s\mathbf{A}\left(\mathbf{A}^T\mathbf{V}_{ss}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{Y}_s - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\det\left(\mathbf{A}^T\mathbf{V}_{ss}\mathbf{A}\right). \tag{1.146}$$

Analogously to the maximum likelihood method, the vector of first derivatives of $\ln L_R$ is determined as well as the expectation value of the $-\ln L_R$ function. The subsequent steps of this procedure are discussed in more detail in the works of Rao and Molina (2015). The asymptotic variance-covariance matrices of the two above estimators have the following form (Rao and Molina, 2015, p. 103):

$$\breve{D}^2_{\xi}\left(\hat{\boldsymbol{\beta}}_{REML}\right) \approx \breve{D}^2_{\xi}\left(\hat{\boldsymbol{\beta}}_{ML}\right) = \left(\mathbf{X}^T_s\mathbf{V}^{-1}_{ss}\mathbf{X}_s\right)^{-1}, \tag{1.147}$$

$$\breve{D}^2_{\xi}\left(\hat{\boldsymbol{\delta}}_{REML}\right) \approx \breve{D}^2_{\xi}\left(\hat{\boldsymbol{\delta}}_{ML}\right) = \mathbf{I}^{-1}(\delta). \quad (1.148)$$

It should further be noted that they are approximately equal to the matrices determined using the maximum likelihood method.

The third, and therefore final, phase of model building is model verification, where we can distinguish tests concerning the distribution of the random component and random effects and the model parameters. The assumption of normality of the random components and random effects for empirical best linear unbiased predictors under the assumption of linear mixed models is important due to:

- parameter estimation by the Gaussian maximum likelihood method (which, however, can be replaced by a non-normality-proof restricted maximum likelihood method – cf. Jiang (1996)),
- testing the significance of the model parameters using classical tests (but these can be replaced by permutation tests, where normality of the components and random effects is not required, and which will be discussed later in this subsection),
- derivation of an approximation of the mean squared error (Kackar and Harville, 1984; Robinson, 1991; Harville and Jeske, 1992),
- for assessing prediction accuracy using most methods except the naive Taylor series expansion method (e.g. Datta and Lahiri, 2000, p. 623), the jackknife method (Jiang et al., 2002, p. 1803), the weighted jackknife method (Chen and Lahiri, 2003, p. 908) and the parametric bootstrap method (Butar and Lahiri, 2003, p. 66).

While for empirical best linear unbiased predictors the assumption of normality is not needed at the stage of deriving the form of the predictor (and proving its ubiasedness), for empirical best predictors it is necessary (at least for the post-transformation values of the random variable under study).

Żądło (2020) points out that under the assumption of a linear mixed model, normality of the distribution of the random effects vectors and random components, and when the **Z** matrix is of full order, the random effects vectors and the random components have multivariate normal distributions if and only if the **Y** vector has a multivariate normal distribution. Hence, accepting the null hypothesis that the vector **Y** (or the random part of the model i.e. $\mathbf{Zv}+\mathbf{e}$) has a multivariate normal distribution is equivalent to accepting the null hypothesis that the vector of random components and the vector of random effects have multivariate normal distributions. The only remaining problem is that classical tests of normality are applied to independent random variables, while the elements of the **Y** vector in the mixed model (and the residuals in the mixed model) are correlated. However, this problem can be solved by applying a transformation of the residuals of the mixed model using the Cholesky decomposition of the inverse of the variance-covariance matrix estimate, as discussed by Jacqmin–Gadda et al. (2006). Classical tests for the normality of the distribution of the random component include the Shapiro and Wilk (1965) and Lilliefors (1967) tests.

Another important element of this stage is the verification of the significance of the model parameters. In the context of the mixed models considered in this monograph, we can distinguish significance tests for fixed effects and for variance components.

In the case of fixed effects tests, the hypothesis being verified is of the form $H_0 : \beta_i = 0$, while the alternative is $H_1 : \beta_i \neq 0$. Among the classical tests for fixed effects, we can mention, among others, the likelihood quotient test, the Wald test and the conditional t-test and, in the case of simultaneous testing of several parameters, the conditional F-test. It should be emphasised that these tests require that the assumption regarding the normality of the distribution of components and random effects is met. An additional condition that must be met for the likelihood ratio test is the nesting of the models. This criterion is met by using the maximum likelihood method in the model estimation process, with the result that the model containing the selected effect and the model not containing it are nested (Biecek, 2012, p. 60). The statistics of this test, discussed in more detail by Pinheiro and Bates (2000), are based on the value of the likelihood function for the restricted model and the more general model, respectively. When the hypothesis being tested is true, this statistic has a χ^2 distribution with a number of degrees of freedom equal to the difference in the number of parameters in the models used to determine it (Pinheiro and Bates, 2000, p. 83). Thus, when testing for the significance of a single parameter, quantiles of the distribution χ^2 with one degree of freedom must be used. An important drawback of this test, pointed out by Pinheiro and Bates (2000), is its "anti-conservativeness". This test is therefore not recommended for testing the significance of fixed effects. As the simulation studies indicate, the actual p-values can be significantly higher than assumed.

Another test considered is the Wald test, also known as the Z-test (Verbeke and Molenberghs, 2000, p. 56). This test should be used when the number of observations is large relative to the number of model parameters. The Wald test is considered for hypotheses for a fixed matrix $\mathbf{K}$ of the form $H_0 : \mathbf{K}\boldsymbol{\beta} = 0$ and $H_1 : \mathbf{K}\boldsymbol{\beta} \neq 0$, obtained by approximating the distribution of the statistic, more extensively presented in the work of Verbeke and Molenberghs (2000), by a distribution χ^2 with number of degrees of freedom equal to the order of the matrix $\mathbf{K}$. In the case where $\mathbf{K}$ is a vector with elements equal to 0 or 1 for j-th selected parameter, the hypothesis under consideration simplifies to the form $H_0 : \beta_j = 0$, and the test and the associated confidence intervals are obtained from an approximation of the corresponding distribution with a standard normal distribution. However, there is a rather significant problem associated with the use of the Wald test to verify the significance of fixed effects. This test is based on standard error estimates of the mean underestimating the actual $\boldsymbol{\beta}$ values. This is due to the non-accounting for the variability implied by the estimation procedure. This problem was considered, among others, by Dempster et al. (1981). A solution to this issue may be the approximation by Student's t-distribution and the use of conditional t-test and F-test statistics, as discussed more extensively by Wolfinger (1993) and Littell et al. (2006). It should be noted that only in certain special cases does the conditional t-test statistic have an exact t-distribution. Similarly, in the case of the second statistic in question, in most cases it will have a distribution only approximating an F distribution with the number of degrees of freedom depending on the order of the **K** matrix (Frątczak, 2012, pp. 412–413).

In addition to classical tests, permutation tests can also be used to test the significance of fixed effects. It should be noted that they can also be used when assumptions regarding the normality of the distribution of effects and random components are not met. This represents a major advantage of permutation tests. The significance verification procedure in this case can be presented as three steps. It should be noted that the subsequent steps will be analogous for both non-random effects and mixed models. In the first step, the value of the statistic, e.g. the reliability function (L_0), is calculated for the model under consideration and the original dataset. The second step involves the B-fold permutation of the elements of the corresponding column of the matrix **X** and the value of the statistic for the considered model and the data, taking the permutation into account: $L_0^{*,b}$ $(b = 1, \ldots, B)$. The number of iterations of B usually depends on the level of significance adopted. The final step of the procedure is to determine the p-value, based on the following formula (Biecek, 2012, p. 22):

$$p = \frac{1 + \#\{b : L_0^{*,b} > L_0\}}{1 + B}, \tag{1.149}$$

so we count the fraction of cases where $L_0^{*,b} > L_0$. It should be added that the likelihood function in the above procedure can also be replaced by other test statistics, e.g. those used in classic tests. The disadvantage of permutation tests is the time required for the whole procedure, which is mainly due to the number of repetitions adopted and the complexity of the test statistic.

In the paper by Krzciuk and Żądło (2014a), a simulation study compared the properties of the above classical tests and permutation tests, including permutation equivalents of classical tests. Analyses were conducted based on data presented in Särndal et al. (1992) concerning Swedish counties. The simulation study was divided into two parts. The first part examined the probabilities of errors of the first type, while the second examined the power of the tests. The study also considered the problem of not meeting the assumptions regarding the normality of the distribution of random effects and random components preventing the use of classical tests, by also including a shifted exponential distribution in their generation. The results obtained in the simulation study indicate good properties of permutation tests based on the likelihood function and conditional t-test statistics.

The null hypothesis, when tested for variance components, is of the form $H_0 : \sigma_v^2 = 0$. The alternative hypothesis, however, following Biecek (2012), can be written as $H_1 : \sigma_v^2 > 0$. Among the classical tests that also find application in testing the significance of variance components are the likelihood ratio test and the Wald test.

The likelihood ratio test statistic, more extensively discussed by Verbecke and Molenberghs (2000), is based on the likelihood function for the model with and without the random effect under consideration, respectively. It should also be noted that this statistic, under some additional conditions, has an asymptotic distribution χ^2 with a number of degrees of freedom equal to the difference in dimensions of the parameter spaces under consideration. Similar to the form of this test for fixed effects, this variant of the test is also counted among the "anti-conservative" tests, as considered in their work by, among others, Stram and Lee (1994).

In the case of the latter test, i.e. the Wald test, the test statistic, following Verbecke and Molenberghs (2000), is in the form of the ratio of the variance component score and the score of its standard error. It should also be added that the distribution of the Wald test statistic for the variance components can be approximated by a normal distribution. It can be added, following Bishop et al. (1975), that for large samples and small deviations from the null hypothesis, both of the above tests will give very similar results. In contrast, the results of analyses by Cox and Hinkley (1974) and McCullagh and Nelder (1989) indicate that for small and medium samples, the likelihood ratio test has better properties.

As in the case of fixed effects, it is also possible to verify the significance of variance components using permutation tests, including permutation versions of classic tests. For tests of variance components the procedure is analogous to that for fixed effects tests. It should be noted that in this case, however, those rows of the **Z** matrix that relate to the random effect under test are permuted.

The paper by Krzciuk and Żądło (2014b) addresses the problem of simulation-based comparison of the properties of classical and permutation tests for variance components, in the context of probabilities of errors of the first and second type. The authors considered the likelihood ratio test and the Wald test, together with their permutation versions and a permutation test based on the likelihood function. The dataset analysed was the data considered in Särndal et al. (1992), concerning Swedish counties. In addition, the analysis addressed the issue of the distribution of effects and random components, relevant for classical tests, by including both a normal distribution and a shifted exponential distribution in the data generation process. This made it possible to show the effect of not satisfying the assumption regarding the normality of the distribution, necessary for the use of classical tests, on the properties of the tests considered. The results obtained suggest good properties of permutation versions of classical tests based on the logarithm of the likelihood function, including in the case of non-fulfilment of the assumption concerning the distribution of random components and effects.

In addition to the issue of the significance of random effects, attention should be paid to the possibility of conducting analyses on, among other things, the existence of correlations of random effects, as well as, in the case of models containing more than one random effect, the possibility of testing the significance of correlations between random effects. The latter issue was proposed by Krzciuk (2018).

When verifying the significance of correlations between random effects, the hypothesis tested is of the form $H_0 : \rho = 0$. The alternative hypothesis can be written as $H_1 : \rho \neq 0$. As for fixed effects and variance components, it is possible to use classical tests, including the likelihood ratio test, but the assumption regarding the distribution of random components and effects must be taken into account. A test based on the parametric bootstrap method, proposed by Krzciuk (2018), also makes it possible to verify the significance of correlations between random effects when the assumption regarding the normality of the distribution is not met. The procedure of this test can be divided into several steps. In the first step, the value of ρ_0 for the original dataset is determined as the correlation coefficient between the random effects scores. The second step involves the B-fold generation of the dataset at the truth of the test hypothesis and estimating from it the value of ρ_{*b}, where $b = 1, 2, \dots, B$. In the last step, the p-value is determined as:

$$p = \frac{1 + \#\{b : |\rho_{*b}| > |\rho_0|\}}{1 + B}. \qquad (1.150)$$

In the paper by Krzciuk (2018), the properties of the two tests above were compared. Analyses were conducted based on the tax revenue data of Swedish counties considered in Särndal et al. (1992). The simulation study addressed both the problem of misspecification of the model in terms of the presence of correlations between random effects and the failure to meet the assumption of normality of the distribution of random effects. Normal distributions, shifted exponential and gamma distributions as well as linking functions (copulas) were used in the data generation process. The analysis was divided into two parts concerning probabilities of type I and II. The results obtained in this paper show the good properties of the proposed test based on the parametric bootstrap method in the context of both type I and II probabilities, as well as the robustness of both analysed tests to the failure of assumptions on the normality of random components in the lack of correlation.

The issue of verifying the significance of both fixed effects and variance components was also addressed in the article by Krzciuk and Żądło (2013). However, the issue was considered in a slightly broader context – longitudinal studies. The analyses used data on own income of Polish poviat budgets from the Local Data Bank (Statistics Poland). The article shows the possibility of also using the discussed tests for longitudinal data using the R language.

1.3. Development of small area estimation

According to Wright (2001), the origins of the survey sampling, of which small area estimation is a branch, date back to the 19th century. We can count the publication of Kiaer (1897) among some of the earliest works on the design-based approach, while the model-based to Cochran (1939). In Poland, however, it has been developed since the 1930s. Key publications from that period include the works of Spława-Neyman (1933) and Piekałkiewicz (1934).

The limited resources allocated to surveys and the associated, also in many cases, inability to increase sample sizes were some of the stimuli for attempts to develop new estimation methods, methods that would allow reliable estimates to be obtained even with small sample sizes. The solution to this problem was to be found in small area estimation methods. The first work in this area was undertaken in the 1970s. Some of the most important publications on small area estimation during this period were, according to Bracha (1996), the work of Gonzalez and Hoza (1978) as well as Purcell and Kish (1979, 1980). The 1980s saw an increase in interest in small area estimation issues and the publication that can be considered a landmark from this period is the work of Särndal (1981).

Numerous scientific conferences and seminars have also contributed much to the development of small area estimation. Among the first would be the international symposium in Ottawa in 1985 (Płatek et al., 1987), as well as the international scientific conferences in Warsaw (Kalton et al., 1993) and Riga, organised in 1992 and 1999, respectively. It should be noted that the Warsaw conference resulted in increased interest in small area estimation in many scientific centres in Poland, including Warsaw, Poznań, Łódź, and Katowice. Among the publications written after the Warsaw conference, we can mention the works: Bracha (1994, 1996), Kordos (1992, 1997, 1999), Gołata (1996), Paradysz (1998), Domański and Pruska (1996, 1997), and Kubacki (1997). Also of great importance was the organisation of the periodic international conference Small Area Estimation. It was held for the first time in 2005 in Jyväskylä and was organised by the University of Jyväskylä, Statistics Finland and the EURAREA Consortium. Each successive edition of this conference is held in a different location in the world, among which can be mentioned: Pisa (Italy), Elche (Spain), Trier (Germany), Maastricht (Netherlands), Shanghai (China), and Maryland (USA). It should be added that the fifth edition taking place in 2014 was held in Poland (in Poznań). Also linked to the Small Area Estimation conference are the meetings of the International Statistical Institiut – ISI Satellite Metting, held in Thailand (2013), Chile (2015), France (2017), Malaysia (2019), and Italy (2021), among others. The last meeting to date was combined with the 2020/2021 edition of the SAE Conference held in Naples, Italy.

Among the more important handbooks and monographs on small area estimation, we can include the works of Rao (2003), Longford (2005), Rao and Molina (2015), Pratesi (2016), and Rahman and Harding (2016). The publications by Rao (2003) and Rao and Molina (2015) address both direct and indirect estimation issues, as well as methods based on small area models, including those on empirical best linear unbiased predictors. Some of the above work was also focused on Bayesian methods. Pratesi (2016), however, provided an overview of small area estimation methods for poverty estimation. Among the issues discussed are temporal-spatial modelling of poverty, estimation of income and inequality distributions. The study also presents examples of the application of small area estimation based on real data. Next authors Rahman and Harding (2016), nevertheless, show the applicability of small area estimation methods to spatial microsimulation modelling, which provide a new approach to creating synthetic spatial microdata. The authors also demonstrate the practical application of the techniques discussed to a range of substantive problems, including how to create models, organise and combine data, and create synthetic microdata. In contrast, Longford's (2005) work addresses the problem of estimation for small areas in the context of the problem of missing data and inference from subdomains poorly represented in the sample. Further works on small area estimation include Dol (1991), Mukhopadhyay (1998), and the National Research Council (1980).

Polish handbooks and monographs on small area estimation include works by Domański and Pruska (2001), Dehnel (2003, 2010), Gołata (2004), Żądło (2008, 2015), Bartosińska (2008), Niemiro and Wesołowski (2010), and Szymkowiak (2020). Domański and Pruska (2001) in their work discussed the basic concepts of both the survey sampling method and small area estimation. The authors also presented the issue of inferring small area characteristics for three sampling schemes – individual unconstrained sampling, stratified sampling and two-stage sampling. The book by Dehnel (2003) deals with the problem of applying small area estimation to assess the economic development of regions. The author's second publication addresses the possibility of using small area estimation to assess micro-enterprise development from an industry-regional perspective. Gołata (2004), however, addresses the issue of the method of estimating the size of unemployment in local terms. The author presented the possibility of using indirect estimation in this issue on the example of the Wielkopolska Voivodeship. Żądło (2008) presented basic issues of the three main approaches in small area estimation – randomised, model-based and model-assisted. The paper discusses both the most important concepts of SAE, as well as selected estimators and predictors. The publication by Żądło (2015) discusses the issues of model-based and model-assisted approaches more extensively. The book focuses a lot on empirical best linear unbiased predictors in single-period and longitudinal studies. The author also presented his own proposal of a super-population model assumed for profiles, taking into account, among other things, the occurrence of correlations in time and space, as well as changes in the population and the affiliation of population elements to subpopulations. Bartosińska (2008) presented the possibility of using indirect estimation methods in representative agricultural research. Paper also addressed the issue of the organisation of representative agricultural research in Poland and the use of alternative sources of information in research conducted worldwide. Niemiro and Wesołowski (2010) consider the use of a hierarchical Bayesian model in the context of a simulation Gibbs sampler study. Kowalczyk (2013) addressed the issue of complex estimation in sample surveys. It should be added that the author considered data from rotational samples in her analyses. Szymkowiak (2020) considered the problem of applying the calibration approach in socio-economic research in his thesis.

European grants have also been important for the development of small area estimation through research projects like EURAREA, SAMPLE, and BIAS. The EURAREA project was carried out for the first time in 2001, and was funded under the 5th EU Framework Programme and by EUROSTAT. The project involved the empirical assessment of commonly used estimation methods in the field of small area estimation and the improvement and extension of these methods. The SAMPLE (Small Area Methods for Poverty and Living Condition Estimates) project was carried out between 2008 and 2011 and was funded by the European Commission under the 7th EU Framework Programme. The main objectives of the project were to develop new indicators and models to better understand the phenomena of inequality and poverty, in particular social exclusion and deprivation. The latter project – BIAS (Bayesian methods for combining multiple Individual and Aggregate data Sources in observational studies) – which ran from 2005 to 2011, was funded by the Economic and Social Research Council's and the National Centre for Research Methods. Its main objective was to develop a methodological framework for dealing with combining data from multiple sources, including combining individual and aggregate data. Mention should also be made of projects such as AMELI and ESSnet-SAE. The analyses carried out in the AMELI project (Advanced Methodology for European Laeken Indicators) included, among other things, studies on data quality and the estimation of small area characteristics and the measurement of change over time. The main objective of the ESSnet-SAE project was to create common procedures and a methodological framework for statistical offices, allowing the development of small area estimation and the sharing of knowledge held by individual offices.

1.4. Applications of small area estimation

The growing demand for information of a local nature, as well as the need for low-cost methods to quickly obtain reliable estimates of subpopulation characteristics, is one of the reasons why small area estimation methods have found and continue to find applications in so many areas. Also, the growing importance of regions – politics, or regional self-government as well as national databases that take into account a very detailed territorial division – has a significant impact on the multitude of areas in which small area estimation approaches can be developed.

In this subsection, selected areas of application of small area estimation methods will be presented together with examples of an economic nature. The following should be mentioned first and foremost: market analyses, labour market analyses including the phenomenon of unemployment, quality of life analyses including the phenomenon of poverty, regional policy, economic aspects of health and environmental policy and agricultural economics.

In the area of market analysis, the approaches used in small area estimation have been applied, among others, in the estimation of house prices. Pereira and Coelho (2013) included both elements of randomised and model-based approaches in their analyses, conducting their considerations using actual data from the Transaction House Prices Survey and the Bank Evaluation of House Prices Survey. Goodman and Thibodeau (1998) addressed the issue of housing market segmentation within metropolitan areas in their research. They used data on single-family home sales transactions in their analyses. The methods of small area estimation can also be used in business analyses. A paper by Nekrasaite-Liega et al. (2011) addressed the problem of estimating corporate income using GREG-type estimators. The analyses were based on data from a quarterly survey of short-term statistics of service enterprises. Dehnel (2018) applied methods from the model-assisted approach to estimating the average revenue of small businesses in the context of assessing the impact of model selection on the quality of estimates.

The analyses used data extracted from administrative records. The market analysis problem was also addressed in the work of Hermalin and Wallace (2001) in the context of estimating the relationship between wages and productivity. The authors used unbalanced panel data on savings and credit institutions in their analyses. Applications of small area estimation in market analyses can also be found in a number of other publications, including Longford (2006), and Domański and Pruska (1997).

Small area estimation methods are also used in labour market research. Molina et al. (2007) addressed the problem of estimating the labour force participation rate using data from the Office for National Statistics of the UK on the UK labour force stock. Longford's (2004) paper showed the feasibility of using the shrinkage estimator to estimate the unemployment rate and the inactivity rate using UK Labour Force Survey data as an example. Ferrante and Pacei (2004) showed the applicability of estimators that are a modification of the Fuller (1990) estimator to estimate the labour force stock. The analyses used data from the Italian Labour Force Survey. The paper by Ręklewski and Śliwicki (2016) considered the problem of estimating the number of economically inactive in the districts of the Kujawsko-Pomorskie Voivodeship. The analyses used data from the Labour Force Survey and the National Census. Analyses related to the labour market in Poland can also be found in the work of Klimanek (2012). The problem addressed in this publication is the estimation of the percentage of unemployed at a lower level of aggregation than presented in Statistics Poland publications using the indirect estimation method. Gołata (2004), nevertheless, concerned indirect estimation of unemployment on the local labour market, estimation of the size of unemployment and estimation of the number of unemployed and employed in poviats of the Wielkopolska Voivodeship. The analyses used data from the micro-census, the Labour Force Survey conducted by Statistics Poland, and registers of the unemployed maintained by labour offices. Applications of small area estimation to labour market issues have also been considered by other authors, e.g. Falorsi et al. (1998), Pfeffermann and Tiller (2006), López-Vizcaíno et al. (2015) or van den Brakel and Krieg (2016).

Small area estimation approaches also allow analyses to assess the quality of life, economic and social situation of the population. Marchetti and Secondi (2017) show that these methods can be used to estimate household consumption expenditure. The authors based their study on information obtained from the Household Budget Survey conducted by ISTAT in Italy. Pratesi and Salvati (2008) used a model-based approach in their analysis of this issue and data obtained from several sources, including the Survey on Life Conditions, the databases of the Istituto Regionale Programmazione Economica and administrative records. Another issue where the methods of small area estimation can be applied is the problem of small area income prediction, as considered in their paper by Fay and Herriot (1979). They used census and household data in their analysis. The methods of small area estimation as well as data from longitudinal surveys are also used in poverty analyses. The paper by Molina and Rao (2010) considers the issue of estimating measures to analyse this phenomenon. The analyses were based on real data from the European Survey on Income and Living Conditions (EUSILC). This problem was also addressed by Diallo and Rao (2018). In his research, Wawrowski (2012) discussed the issue of estimating the poverty risk rate using direct and indirect estimation methods on the example of the districts of the Wielkopolska Voivodeship. The data on which the study was based came from the Statistics Poland's Household Budget Survey and the National Census. The problem of poverty estimation was also addressed by, among others, Graf et al. (2018), and Tanton et al. (2011). The work of Kriegler and Berk (2010) demonstrates the feasibility of using small area estimation methods to estimate the number of homeless people in Los Angeles. The model-based approach can also be used to improve the accuracy of prediction of, for example, the number of crimes per 100 inhabitants or the number of drug-related crimes, as addressed by van den Brakel et al. (2016). This research was based on data from the Police Administration of Reported Offences, the Dutch National Safety Monitor and the Integrated Safety Monitor.

The model-based approach is also applicable in regional analyses. The paper by Jędrzejczak and Kubacki (2017) considers the application of the EBLUP and the multivariate model of Rao and Yu (1994) to predict per capita income and expenditure in regions in Poland. The study used data from administrative records as well as from the Household Budget Survey. The approaches used in small area estimation are also applied in the assessment of the economic development of regions, which was considered by Dehnel (2003). In her study, the author used data from four sources: the monthly report on economic activity, the file of economic entities, compiled based on the REGON system, the Local Data Bank of the Statistics Poland, and the National Census. Jhun et al. (2003), however, considered the problem of predicting the Cobb–Douglas function for panel data. The study used panel data from Munnell (1990).

Small area estimation methods also allow analyses to be carried out to investigate the influence of specific factors on the incidence of selected diseases. In the work of Lawson et al. (2012), analyses were carried out on the effect of PM2.5 concentrations on the incidence of asthma, based on data considered by Fuentes et al. (2006) and Choi et al. (2009). The issue of the influence of genetic, environmental and age factors on the incidence of respiratory diseases was also considered by Torabi and Shokoohi (2015). In analyses of the economic aspects of health policy, approaches used in small area estimation can be used to estimate the number of people with disabilities, or the use of health services. Such studies are conducted in the United

States under federal programmes. Analyses of people with disabilities are presented, among others, in You et al. (2014).

In agricultural analyses, methods of small area estimation are used, among other things, to predict the crop area in regions, as presented in the article by Battese et al. (1988). In this work, sample survey data and satellite data were used. The issue was also considered in the work of Militino et al. (2006) in the context of estimating the total area of olive trees in a region in Spain. Considerations on the application of small area estimation to estimate the percentage of indebted farms are presented in Chandra et al. (2018). The authors used data from the National Sample Survey Office on agricultural areas of the state of Bihar in India. Applications of small area estimation to agricultural economics problems were also considered in the papers of Lohr and Prasad (2003), Torabi and Rao (2010), Rivest et al. (2016), and Fabrizi et al. (2014).

The methods of small area estimation are also used in analyses that may have implications for environmental policy. These issues have been addressed in the work of Opsomer et al. (2008), and Petrucci and Salvati (2006). The authors of the first paper applied the proposed estimation methods to estimate the average acid neutralising capacity of lakes in the northern US states. In the work of Petrucci and Salvati (2006), a simulation study was conducted to compare the properties of the selected estimators and the authors' proposed spatial EBLUP. Floodplain erosion data for land located near the Rathbun Lake Watershed in Iowa were used in the conducted considerations.

1.5. Summary

This chapter discussed the theoretical foundations of small area estimation. Subsection 1.1 is focused on the main approaches in small area estimation – the randomised, model-based and model-assisted approaches. For each of these, the basic concepts and selected estimators or predictors were presented. For selected predictors in the model-based approach, generalisations to longitudinal data were shown.

In subsection 1.2, the problem of building a superpopulation model was discussed in more detail. The first part of the subsection presented a classification of overpopulation models with the greatest emphasis on linear mixed models. In the next part, special cases of this class of models were discussed with a distinction between models with uncorrelated and correlated random effects. Special attention should be paid to the proposed special cases of general mixed models with correlated random effects vectors in longitudinal studies (cf. Krzciuk, 2020). The remaining parts of this subsection dealt with the different steps of the process of building a superpopulation model, and therefore its specification, estimation and verification. For each of these stages, most attention was paid to the methods used for linear mixed models. In the section on the verification of the superpopulation model, a proposal was also made for a test based on the parametric bootstrap method, which makes it possible to verify the presence of dependencies between random effects. Attention was also drawn to the applicability of permutation equivalents of classical tests and their good properties based on simulation analyses conducted in this area.

Subchapter 1.3 was focused on the development of small area estimation, both worldwide and in Poland. It discusses the reasons for the growing interest in small area estimation methods. Selected conferences, projects and publications of key importance for the development of small area estimation were also presented.

In subsection 1.4, selected areas of application of small area estimation methods were discussed. Among them are analyses of, for example, the market, the labour market, poverty, and regional policy. It should be added that numerous examples of analyses of an economic nature were presented for each of the areas mentioned.

The author's theoretical proposals presented in this chapter include applications of LMMs with correlated random effects vectors in small area estimation, including longitudinal studies. In addition, generalisations of selected predictors to the case of cross-sectional-temporal data were presented. The chapter presented the possibility of using permutation tests in the verification of the significance of the parameters of the proposed models, as well as the author's test based on the parametric bootstrap method to verify the presence of correlations between random effects.

Chapter 2

Cross-sectional and longitudinal economic surveys

This chapter will be focused on the issues of cross-sectional surveys and repeated surveys over time. In the following subsections, the essence of cross-sectional and longitudinal surveys will be presented. A classification of repeated surveys over time with economic examples will be presented. The advantages and disadvantages of longitudinal surveys will also be discussed.

2.1. Research conducted during one period

This subsection will discuss the concept of a statistical survey as well as the steps involved in conducting one. A classification of statistical surveys and examples of their applications will also be presented. A statistical survey should cover a specific statistical population. Its aim is to determine the regularities occurring in the analysed community, based on the features that characterise the units comprising it (Sobczyk, 2004, p. 16).

Among statistical surveys, we can distinguish between full surveys, also called complete or exhaustive surveys, which cover all units belonging to a community, and partial (incomplete) surveys, which involve only a certain subset of elements – a sample. In the class of full surveys, we distinguish between statistical censuses, current registration and statistical reporting. Examples of full surveys include The General Agricultural Census and the National Population and Housing Census conducted by Statistics Poland, and current registrations of births, deaths as well as marriages. Statistical reporting includes, inter alia, reports of enterprises or companies (Starzyńska, 2005, pp. 23–24). Based on current registrations, administrative registers such as REGON, PESEL and ZUS are created.

Within the framework of incomplete research, we can distinguish surveys, and monographic and representative research (Sobczyk, 2004, p. 17). Following Mazurek-Łopacińska (2002) and Kędzior (2005), we can classify survey research by a number of criteria, e.g: confidentiality (open and anonymous research), the technique of filling in the questionnaire (traditional or electronic), the method of delivering the questionnaire to the respondent (e.g. by post, electronic, telephone, distributed), the participation of the interviewer (with or without the participation of the interviewer). Survey research in Poland is conducted by such institutions as TNS Polska (formed by the merger of TNS OBOP and TNS Pentor), the Centre for Public Opinion Research, and the Resort for Public Opinion Research. Monographic surveys are, as Sztumski (2004) notes, surveys characterised by the focus of the analysis of a specific object – a statistical unit or a small group of them, a high level of detail in the analysis of the researched phenomenon and the interdisciplinarity of the conducted research process. Sobczyk (2004) also points out that the unit which is the subject of the monographic study should be typical for the population from which it was selected or, on the contrary, a unit setting the direction of development in the studied population. Sample surveys are surveys conducted on the basis of sample data. The analyses conducted for this type of research use the approaches discussed in the first chapter of this book.

It should be added that full surveys, especially in the case of very large populations, are associated with high costs of carrying them out as well as the long time needed for their implementation. In such cases, the decision is most often made to conduct a partial survey. According to Sobczyk (2004), such a choice is also made in the case of surveys that are destructive in nature.

Zeliaś et al. (2002) distinguish four steps in conducting a statistical survey. The first is survey preparation. This should include defining the purpose and method of the survey. In this step, the population and statistical unit as well as the characteristics to be analysed should also be defined. The second phase is statistical observation. According to Sobczyk (2004), the result of this stage is to obtain a set of data referred to as statistical material. The next step in the process of conducting a statistical survey is to process the collected material. As a result of these activities, statistical series are obtained for the surveyed variables. In addition, the collected data can also be presented in the form of graphs at this stage (Sobczyk, 2004, pp. 21–23). The last stage involves statistical analysis. Most often, in this phase, a statistical description of the surveyed population or sample is made or a statistical inference is made, allowing the results obtained on the basis of the sample to be generalised to the population. This phase is completed by characterising the phenomenon under study and drawing conclusions based on the results obtained. The above division of the process of conducting a statistical survey into phases also applies to repeated surveys over time, which will be discussed in more detail in the following subsections of the chapter.

2.2. Essence of longitudinal surveys

There has been an increase in interest in longitudinal studies, according to Nathan (2009), in recent decades. Previously, analyses in the social sciences were mainly based on cross-sectional survey data. The reason why time-repeated studies were not often used was the cost of their implementation as well as the operational and methodological complexity. In the earliest surveys, the time over which an individual is observed was usually quite short, limiting the possibility of using the data obtained at the micro or individual level. Initially, therefore, the repetition of surveys over time served to increase efficiency, the effectiveness of cross-sectional surveys and estimates of change at the macro level, rather than to analyse flows or gross changes and therefore changes at the level of individuals. In recent decades, however, efforts have begun to enable the introduction of advanced methods for the long-term analysis of social and economic processes, resulting in a significant increase in the importance of longitudinal data as a basis for empirical research in the social sciences, which includes economics (Nathan, 2009, p. 315).

Fitzmaurice et al. (2004) identify the characterisation of the change in the respondent's response and, consequently, the selected characteristic over time, as the main objective of longitudinal surveys. It is also possible to identify factors influencing these changes. The work of Duncan and Kalton (1987) sets out the aims of repeated surveys over time in somewhat more detail. The authors listed seven basic objectives. Kalton (2009) points out that it is possible to classify the above objectives and to assign types of research that implement them. The types of research along with their characteristics and examples of application will be discussed further in the next subsection of this book.

The first set of objectives identified by Kalton (2009) includes estimating population characteristics at different times or periods if changes can be treated as insignificant. Another objective is to enable the determination of the averaged estimates of population characteristics for several periods. A third objective belonging to this group is to assess the net change in the characteristics under study, and thus to assess the dynamics of change in characteristics between periods at a higher level of aggregation. It should be noted that the above objectives do not impose any conditions on the relationship between samples at different points in time. They can therefore be met in surveys with complete rotation or surveys that minimise the overlap of samples over time, including rotation panels. In particular, these objectives can be met by collecting cross-sectional survey data from different moments in time (Kalton, 2009, p. 90).

The next group is also made up of three objectives. It is designed to assess the various components of individual change, including the gross change, average change at unit level, and volatility (variance) for each unit. This class also includes the aggregation of data for individual units over time, e.g. quarterly to annual data. Kalton (2009) cites as an objective belonging to this class the measurement of the frequency of occurrence of the phenomenon under study or its duration over the period under study. It should be added that this class of objectives can only be realised through panel surveys, as it requires having data from the same subset of units (Kalton, 2009, p. 90).

The last objective mentioned in Duncan and Kalton's (1987) work, i.e. the accumulation of samples over time, particularly when studying rare phenomena, was assigned by Kalton (2009) to a separate class. It should be noted that where a rare feature relates to an event, such as divorce, this objective can be met using any type of longitudinal survey. However, if it is of a fixed nature, such as racial group membership, the objective can only be met if the sample is replaced in subsequent periods (Kalton, 2009, pp. 90–91).

Steel and McLaren (2009) also highlight key elements that should be considered when designing longitudinal studies. The authors mention, among others, the frequency of sampling and the schema for including new units in the survey. The decision regarding the intervals between successive waves of the survey is primarily driven by the purpose of the analysis and the characteristics under consideration. Most often, successive survey rounds are conducted annually, quarterly or monthly. Also related to the issue of whether or not to include overlapping samples is the determination of how long individual units remain in the sample. The problem of designing longitudinal surveys has been considered by Binder and Hildegrou (1988), Kalton and Citro (1993) and Steel (2004), among others.

2.3. Types of longitudinal surveys

In this subsection, a classification of longitudinal surveys will be introduced. For each type of research, the manner in which it is conducted will be discussed and presented in the form of a scheme. The description of research will also be supplemented by selected examples of its implementation in Poland and worldwide, which are of an economic nature.

2.3.1. Panel studies

The first type of survey to be discussed is the panel. In panel studies, the same individuals drawn from the population are analysed repeatedly, at different times. In a schematic way, assuming for simplicity that the composition of the population does not change over time, this type of survey is shown in Figure 2.1. The grey colour indicates the population elements that have been drawn for the sample. The following lines show the composition of the sample at specific points in time. The origins of the use of panel data in research date back to the 1940s. At that time, Paul F. Lazarsfeld (Lazarsfeld and Fiske, 1938; Lazarsfeld, 1940) attempted to introduce this method into market research and public opinion analysis (Andreß et al., 2013, p. 1). One of the earliest panel studies, known as the Erie County study, concerned the analysis of voting behaviour during the 1940 presidential campaign. It was conducted by Columbia University's Bureau of Applied Social Research under the direction of Lazarsfeld (Lazarsfeld et al., 1944).

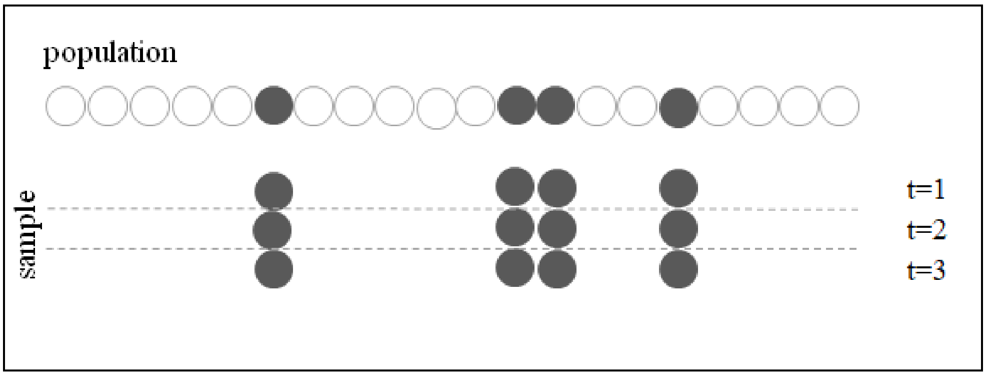


Figure 2.1. Panel study scheme

Source: Own elaboration.

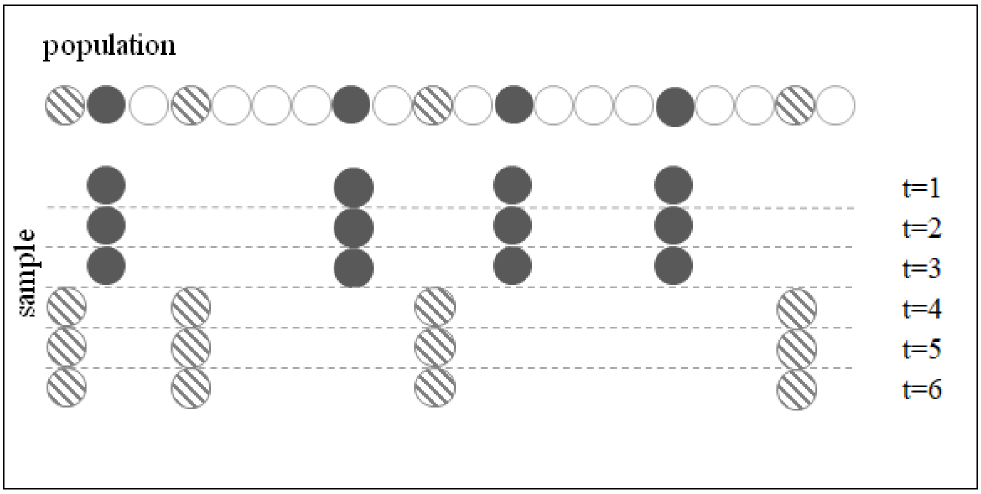


Figure 2.2. Scheme of a panel study without overlapping samples

Source: Own elaboration.

It should be noted that one can find a division of panel studies into balanced or unbalanced in the literature. In the balanced panel, each unit is surveyed exactly the same number of times, while in the unbalanced panel, the number of observations for each unit is not the same. Thus, an unbalanced panel may occur when some respondents do not participate in all measurement periods or leave the panel before completing it (Andreß et al., 2013, p. 62).

In Kowalczyk's (2001) study, surveys consisting of several panels were also distinguished, among which one can distinguish between repetitive panel surveys without overlapping and with overlapping samples, which are shown schematically in Figures 2.2 and 2.3, respectively. To distinguish between the elements drawn in the first part of the survey, they are marked in grey, while the second, with a pattern of white and grey stripes.

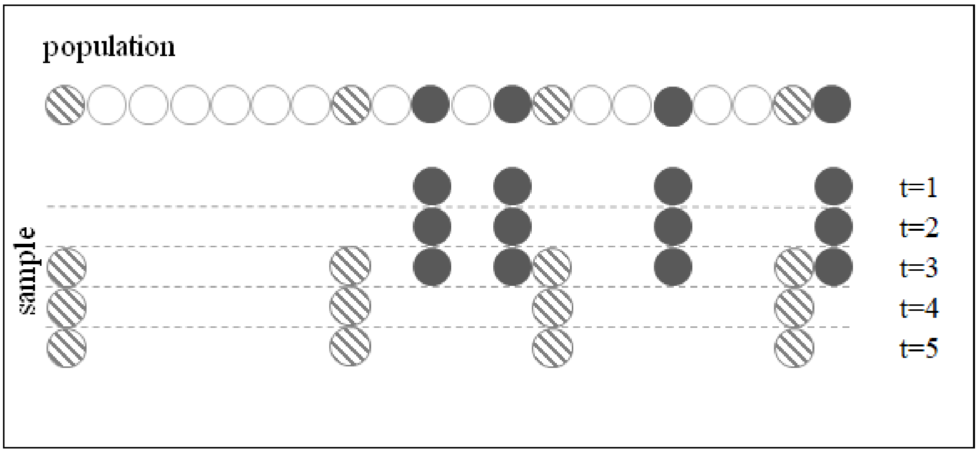


Figure 2.3. Scheme of a panel study with sample overlap

Source: Own elaboration.

Here, as in the case of subsequent schemes, the assumption is made that the composition of the population is invariant over time. The author notes that formally, repeated panel surveys with overlapping samples are often categorised as surveys with partial rotation, which will be discussed later in this paper. However, in contrast, they are mainly used for analyses of the duration of a given condition, e.g. use of social assistance benefits or being unemployed. They are also characterised by a longer implementation period and fewer separate panels (Kowalczyk, 2001, p. 33).

Among panel studies, a group of cohort studies can also be distinguished. According to Balicki (1986), a cohort is a group of individuals who experienced the same event at the same time and place. In economic research, this concept can also be applied to other statistical units, among others, enterprises. Cohort studies include the National Longitudinal Surveys, funded by the U.S. Bureau of Labor Statistics, the Survey of Health, Ageing and Retirement in Europe (SHARE Project) and the English Longitudinal Study of Ageing funded by the National Institute on Aging in the US, and the Health and Retirement Study. In Poland, the issue of cohort studies has also been addressed in the works of Gołata (1995), Balicki (1997), Jackowska (2015), Markowicz (2012, 2016), and Mikulec (2017), among others.

In practice, due to the difficulties and certain drawbacks of this type of research, which will be discussed in more detail later in this chapter, panel surveys more often than not take the form of an unbalanced panel. Their methodology is also subject to certain modifications, such as supplementing and expanding the initial sample, among others. They therefore become closer to rotational panels or surveys that are a combination with partially rotated surveys.

An example of a panel survey conducted in Poland is the POLPAN Polish Panel Survey. It has been conducted since 1988, with successive editions conducted every five years. The POLPAN survey is conducted by a research team working at the Institute of Philosophy and Sociology of the Polish Academy of Sciences, under the direction of Kazimierz M. Słomczyński (Słomczyński with the team, 2014, p. 11). As pointed out by Kiersztyn et al. (2017), POLPAN is one of the longest-running panel surveys in Central and Eastern Europe. The questionnaire used in the survey includes questions on, among other things, working life, education and views on changes in the political and economic system. The main purpose of the POLPAN survey is both to provide a comprehensive characterisation of the social structure in Poland and to enable analyses from a dynamic perspective. In particular, the dynamic approach allows for analyses concerning the adaptation of Polish society to the economic and political changes that took place after 1988 (Słomczyński with the team, 2014, p. 2). In the first edition of POLPAN, in the process of drawing units for the survey, the data available to the Public Opinion Research Centre from the "micro-census" conducted in 1986 were used as the sampling frame. The sample in this survey was furthermore a two-stage sample, where the units drawn in the first stage were census districts and in the second stage were households. Starting from the third edition, the survey was supplemented by a sample of young people aged 21–30, drawn on the basis of the PESEL. (Słomczyński with the team, 2014, p. 5; POLPAN methodology).

Research of a panel nature can also include the Social Diagnosis project, which began in 2000 and was initiated by Wiesław Łagodziński. The main objective of the project carried out by the Council for Social Monitoring is to obtain information on the most important aspects of the life of households and the people who make them up. The survey covers both economic and non-economic aspects related to living conditions, its quality and style, as well as the demographic and social structure of households. Data are obtained, among others, on income situation, material prosperity, housing conditions, participation in culture and recreation, use of services of the health care system and social assistance received by the household. The survey also provides information on the subjective assessment of the material standard of living, the value system, attitudes and social behaviour of the individual members of these households. The survey is conducted at two-year intervals, with the exception of 2003, when the interval was three years from the previous edition (Czapiński and Panek, Eds., 2015, pp. 13–14). It should be noted that two questionnaires are used in the Social Diagnosis. The first one concerns the household and is filled in by the interviewer, the second one is filled in individually by all household members aged 16 and over. The survey is supervised by the Office for Statistical Research and Analysis of the Polish Statistical Association. From the second edition onwards, the survey includes persons who arrived in the surveyed households and households formed by splitting the initial sample (Czapiński and Panek, Eds., 2015, pp. 25–26). It should be noted that the initial sample of households was a two-stage sample, where the first-stage sampling units were statistical districts or census tracts and the second-stage sampling units were dwellings. In addition, prior to drawing the units for the survey, a stratification by place of residence – provinces and class of locality – was made. The same number of households was drawn from each voivodeship, and parameter estimates for Poland as a whole were obtained as weighted averages of the data for the voivodeships (Czapiński and Panek, Eds., 2000, p. 9).

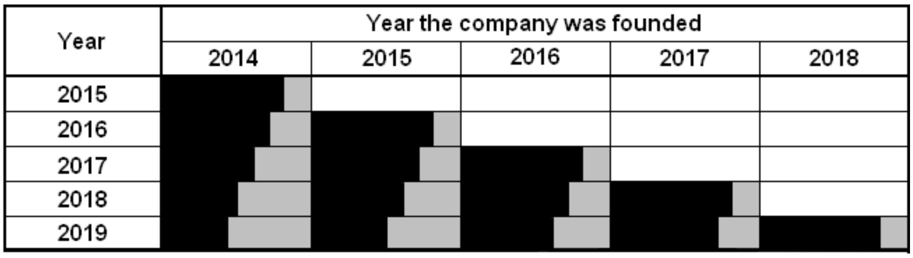


Figure 2.4. Scheme of the company panel survey

Source: Own elaboration based on Statistics Poland (2015, p. 15).

The panel survey of enterprises carried out by Statistics Poland since 2002 can be classified as a survey conducted in the form of several parallel panels, shifted by one year. It produces data on the conditions of establishment and development, the current situation, as well as the survival rate of enterprises in the first years after registration. The condition of enterprises is determined on the basis of, among other things, the extent of their activities, financial result, number of employees, and sources of financing. The survey is conducted on a five-year cycle. The enterprises surveyed in the first period include those that were established (registered) in the year preceding the survey. In the following four years, enterprises are included in the survey if they are still operational and active at the time of the survey (permanent or seasonal). After this period, enterprises are considered stable and excluded from the panel (Statistics Poland, 2015, pp. 13–15). A scheme of the survey is presented in Figure 2.4. Surveyed (active) units are marked in black, while inactive units, i.e. not participating in the survey, are marked in grey. It should be noted that until 2005, only micro-enterprises (employing less than 10 persons) were included in the survey. In the later surveys, small enterprises with up to 49 employees were also included. In addition, the survey uses a sampling scheme based on sampling without replacement from strata with proportional allocation. The strata were distinguished on the basis of the type of business conducted, legal form and size of the entity (Statistics Poland, 2012, pp. 16–17). The Survey of Income and Program Participation (SIPP) and the Survey of Labour and Income Dynamics (SLID) conducted by the United States Census Bureau and Statistics Canada, respectively, are also examples of surveys in the form of several parallel panels staggered relative to each other.

One example of a panel survey conducted internationally is the British Household Panel Survey (BHPS). This was conducted for 18 years until 2009, when it was included in UK Households: a Longitudinal Study. The BHPS survey was funded by the Economic and Social Research Council. The sampling used a cluster random sampling design and the postal address register was used as the sampling frame. Before the introduction of the youth questionnaire, in 1994, respondents had to be at least 16 years old. In addition to the individual questionnaire, a household questionnaire was also included in this study. In 1999 and 2001, the existing sample was extended to include households from Wales and Scotland as well as Northern Ireland, respectively. The BHPS questionnaire primarily covered issues such as household demographic composition, housing, education, health and medical care, labour market, income and benefits (Taylor et al., Eds., 2018).

The Panel Study of Income Dynamics (PSID) is a panel study that has been conducted in the United States for nearly 50 years. The study was conceived by President Lyndon Johnson and the results were intended to help fight poverty. The 1968 baseline sample was drawn from two independent samples: a sample of low-income families from the Survey of Economic Opportunity and a sample of families developed by the Research Center at the University of Michigan. As in the other surveys, the sample was supplemented with new members of the drawn households and persons co-forming new households with those drawn. In addition, attempts were also made to supplement the baseline sample with a sample of immigrant families, while in 1997 it had to be reduced due to the intensive natural increase in its size generated by the division of households. In its initial editions, the survey focused on employment, income and household demographics. In later years, however, it was expanded to include, among other things, questions on health, religion, use of computers, Internet and other media, and expenditure on education and medical care (Beaule et al., 2017, pp. 10–20).

Among the household panel surveys conducted worldwide, we can also mention, among others, in Europe, the European Community Household Panel, the Panel Study on Belgian Households, the Swedish Panel Study, and the German Socio-Economic Panel, and in Asia, the Japan Household Panel Survey. In addition, we can also include the National Educational Panel

Study conducted by the Leibniz Institute for Educational Trajectories as well as the Medical Expenditures Panel Survey conducted in the United States.

2.3.2. Repeated time surveys with partial rotation

Surveys with partial rotation, also called rotational panel surveys or rotational surveys, are also an important group of surveys that are repeated over time. In this type of survey, part of the sample is replaced by newly added items after a certain period of time. An example of a rotational survey scheme is shown in Figure 2.5. The units drawn in the first period are marked in grey. In subsequent periods, half of the elements participating in the survey in the previous period are replaced. New units drawn in the second and third periods are marked with a grey striped pattern and a white and grey grid, respectively. Rotational panel surveys avoid many of the problems that can arise with zero-rotation surveys. They can also be considered as a certain compromise between panel and complete rotation surveys, which will be discussed in more detail later in the monograph.

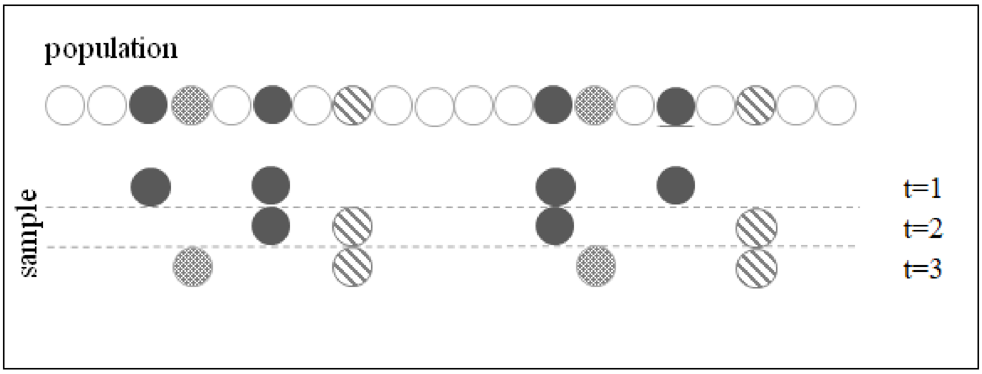
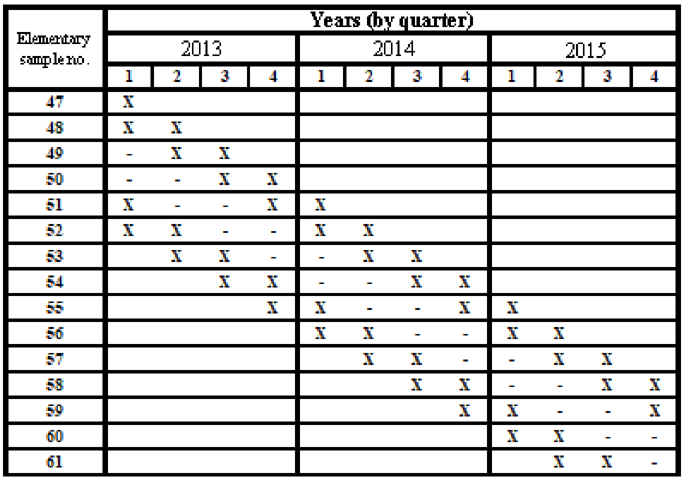


Figure 2.5. Partial rotation survey scheme

Source: Own elaboration.

An example of a survey with partial rotation carried out in Poland is the Labour Force Survey (LFS). It has been carried out quarterly since 1992 by Statistics Poland. The main aim of the LFS is to obtain information on the economic activity of the population, the phenomenon of unemployment and professional inactivity and its causes. The survey covers members of households living in the drawn dwellings aged 15 and over (Statistics Poland, 2013, pp. 11–12). The sample in the LFS is a two-stage sample. The first-stage sampling units are statistical districts or, in the case of rural areas, census tracts, while the second-stage units are dwellings. The information necessary to draw the sample is obtained from the Social Survey Operator, which contains both the list of territorial statistical units and the addresses of dwellings (Statistics Poland, 2017a, p. 13). The selection of the LFS quarterly samples follows the so-called rotation scheme. An example of an excerpt from the rotation chart is shown in Figure 2.6. The quarterly sample consists of four elementary samples each time, divided into thirteen weekly samples. This maintains the continuity of the survey. A partial exchange of elementary samples is carried out each quarter. Two samples surveyed in the previous quarter, plus one newly introduced sample and one sample not surveyed in the previous period, but introduced into the survey exactly one year before. It should further be noted that the elementary samples are drawn independently. It follows from the above scheme that each elementary sample takes part in the survey for four quarters, with a six-month break after two quarters, and thus according to the so-called 2-(2)-2 rule (two quarters in the survey, another two breaks, two quarters in the survey) (Statistics Poland, 2017a, pp. 12–13).



Elementary sample no.	Years (by quarter)											
	2013				2014				2015			
	1	2	3	4	1	2	3	4	1	2	3	4
47	X											
48	X	X										
49	-	X	X									
50	-	-	X	X								
51	X	-	-	X	X							
52	X	X	-	-	X	X						
53		X	X	-	-	X	X					
54			X	X	-	-	X	X				
55				X	X	-	-	X	X			
56					X	X	-	-	X	X		
57						X	X	-	-	X	X	
58							X	X	-	-	X	X
59								X	X	-	-	X
60									X	X	-	-
61										X	X	-

Figure 2.6. Labour Force Survey scheme

Source: Own elaboration based on Statistics Poland (2013, p. 10).

Repeated surveys over time with partial rotation also include the Labour Force Survey (LFS), conducted by the Australian Bureau of Statistics since 1960. Initially, this survey was conducted quarterly and then, from February 1978, monthly. The main purpose of the survey is to provide information on the labour market activity of the Australian population aged 15 years and over. For those who are employed, data are collected on, among other things, their occupation, working hours, and employment status. When the person surveyed is unemployed, the information extracted includes whether they are seeking employment, the length of time they have been unemployed, and the occupation in which they last worked. The questionnaire also includes questions on age, marital status, and education. It should be added that the LFS sample consists of three segments. The first is private houses and flats. The second is hotels, hospitals, homes for the elderly, universities, boarding houses, etc., and the third is flats and houses of people belonging to the Aboriginal and Torres Strait Islander communities. Among others, data from the Geocoded National Address File (G-NAF) and the Australian Statistical Geography Standard (ASGS) are used as the sampling frame in the LFS. A multistage sampling scheme is used in the selection of the sample. Analysing the unit rotation scheme in the Labour Force Survey, it can be seen that it is assumed that the sample is divided into eight sub-samples (or rotation groups). Every month, one eighth of the sample and therefore one subsample is replaced by a new one. The monthly replacement of subsamples ensures that no dwelling remains in the sample for more than eight months. It is also important to note the changes in the way the LFS is conducted. Until August 1996, respondents were interviewed in person. After this period, the option of conducting the survey by telephone or online began to be introduced as well. Rotational panel surveys conducted worldwide also include the Swedish Labor Force, the Labour Force Survey UK or the US Current Population Survey.

2.3.3. Multi-period surveys with complete rotation

A separate type of longitudinal surveys are repeat surveys with complete rotation. In these surveys, a new sample is analysed in each period, as schematically illustrated in Figure 2.7. The units to be surveyed in successive periods are marked with grey, a white-grey striped pattern and a white-grey grid, respectively.

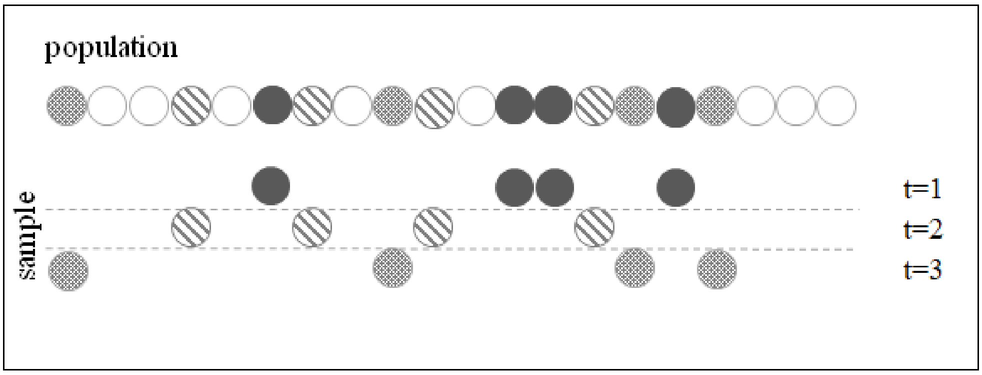


Figure 2.7. Scheme of the survey with complete rotation

Source: Own elaboration.

In Poland, the complete rotation method is used in the household budget survey conducted by Statistics Poland. In this survey, a monthly family exchange period is currently used. During this period, households record their outgoings and incomes in notebooks specially prepared for this purpose – budget books. The survey produces information on demographic and social structure, durable goods owned by households, housing conditions, and material status (Statistics Poland, 2011, pp. 11–15). The sample for this survey is selected using a two-stage sampling scheme, in which the sampling frame is based on the list of statistical regions developed for the

National Census and the register of the territorial division of the country – TERYT system. The first-stage sampling units are statistical regions or groups of regions referred to as field survey points. Second-degree units are individual dwellings. It should be noted that the first-degree units also include stratification by class of locality (Statistics Poland, 2011, pp. 13–14; 2017b, pp. 15–16).

This method is also used in the study, "Tourist activity of Polish residents on tourist trips" (Łaciak, 2013). It is conducted on behalf of the Tourism Department of the Ministry of Sport and Tourism. The main objective of this survey is to obtain information on the types of trips taken by Poles, the directions of trips, both domestic and foreign, and the expenses incurred in connection with trips. The questionnaire also includes questions on the time of travel, purpose, method of organisation, accommodation used, and means of transport used (Łaciak, 2013, pp. 13–14). It should be added that the sample drawn for this study is random-quota. This is to ensure both a random selection of survey locations but also that the sample is in line with the structure of the surveyed population in terms of age and gender. The selection procedure can therefore be divided into two stages. In the first, 200 addresses – starting points – are drawn from the Statistics Polands address database. In the second step, which takes place during the survey, five interviews are carried out from one starting point and the respondents are selected by quota for the characteristics mentioned above. The sampling also takes into account the stratification by province and class of locality. The number of starting points for a stratum is proportional to the number of population meeting the age criterion in this survey (Łaciak, 2013, p. 14).

Surveys with total sample rotation conducted worldwide include The British Social Attitudes Survey conducted by NatCen Social Research. This survey has been conducted since 1983. The main aim of the survey is to obtain information about people's social, political and moral attitudes and the changes in them. The postcode database – Postcode Address File – and a list of addresses from post offices are used as the sampling frame in this survey. It should be added that the sample drawn in The British Social Attitudes Survey is a multi-stage sample. The next steps of sampling individuals are the selection of sectors defined by postcode, addresses and finally respondents (Britsish Social Attitiudes Survey, 2016).

This method was also used in the National Health Interview Survey conducted since 1957 by the National Center for Health Statistics. This survey provides information to monitor the health status of the US population, access to health care and progress towards health policy goals. The information extracted from this survey is used by the Department of Health and Human Services as well as public health research organisations, among others (National Health Interview Survey, 2019). Surveys with complete rotation can also include those conducted by the

CSO on land use, sown area and livestock. Also, some editions of the General Household Survey and the US Health Interview Survey were conducted as rotational surveys.

In practice, surveys that are a combination of panel surveys and surveys with rotation, complete or partial, are also used. Figures 2.8 and 2.9 show a schematic of how this type of survey is carried out. In both cases, half of the elements drawn in the first period (marked in grey) constitute the part that is a panel. In the survey presented in Figure 2.8, the remaining elements are subject to complete rotation, while in Figure 2.9 only to partial rotation.

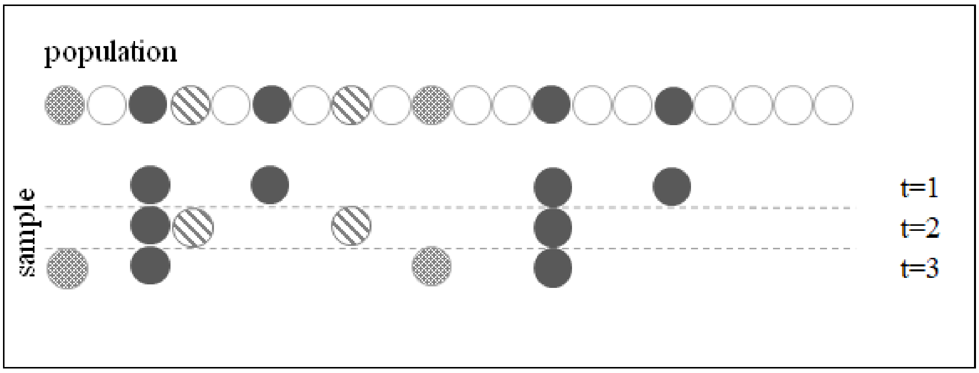


Figure 2.8. Scheme of combinations of panel survey with complete rotation

Source: Own elaboration.

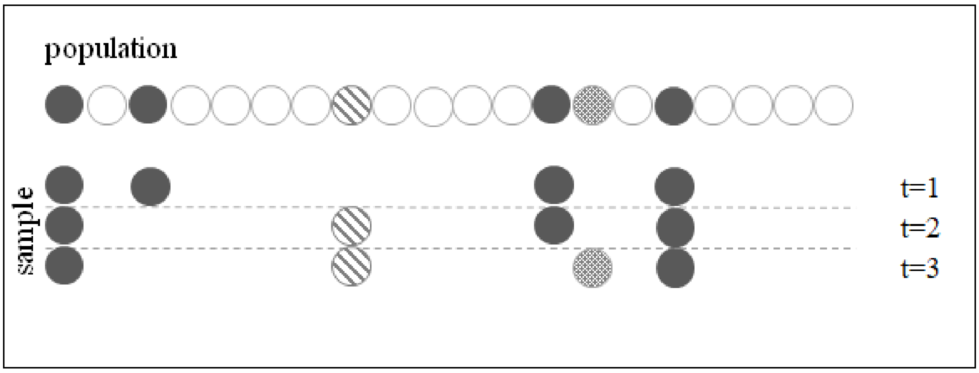


Figure 2.9. Scheme of combination of panel survey with partial rotation

Source: Own elaboration.

2.4. Advantages and disadvantages of longitudinal surveys

When addressing the issue of conducting repeated surveys over time, it is also important to mention the advantages of such surveys as well as their limitations. We should also consider what benefits or difficulties the use of such acquired data will entail.

2.4.1. Advantages of time-repeated surveys

This subsection presents the advantages of longitudinal surveys, taking into account the classification presented in the previous section. It will also discuss the advantages of being able to use data obtained from not one but many periods.

Trivellato (1999) emphasises that panel data make it possible to identify both permanent and temporary characteristics of the phenomenon under study. He also mentions that panel surveys allow analysis of intergenerational behavioural patterns, such as the phenomenon of generational poverty. These patterns are difficult to capture even in retrospective cross-sectional studies. There is also a significant advantage in being able to make more accurate forecasts for individual units by combining data rather than determining individual forecasts using only data for that unit. Importantly, panel data provides the opportunity to learn about an individual's behaviour by observing others whose behaviour is similar. This provides a more accurate description of an individual's behaviour. This issue was considered, for example, by Hsiao et al. (1989; 1993).

Panel data are also important in analyses of micro-dynamic behaviour and micro-social change. They allow the use of models appropriate for these analyses, which take into account the order of events and aim to capture the dynamic relationships between events and behaviour. Cross-sectional-temporal data thus make it possible, among other things, to analyse demographic processes from the point of view of both the determinants and the choices that give rise to certain behaviours, as considered by Courgeau and Lelievre (1988), among others. They also provide a micro basis for analysing aggregate data. This is important if the micro-units are heterogeneous and the time series properties for the aggregate data may differ significantly from the disaggregated data (Hsiao, 2007, p. 5). This problem has been considered by, among others, Granger (1990), Lewbel (1994), and Pesaran (2003). As Hsiao (2007) points out, panel data, due to the fact that they contain observations in the form of time series for many individual units, are ideal for studying issues of homogeneity and heterogeneity. Panel studies can not only be a source of data for analyses of changes in the phenomenon under study over time at the level of individuals, but they can also be used for analyses of dependency over time. An example of this is the analysis given by Ashenfelter and Solon (1982) of the impact of work experience in the labour market on earning capacity in later years, and participation in various government programmes on later economic status. This is very relevant to research of an economic nature because – as Nerlove (2002) points out – economic phenomena are inherently dynamic, and consequently most econometrically interesting correlations are directly or indirectly dynamic.

Panel data also capture the complexity of human behaviour better than with cross-sectional or time series data. Among other things, they allow more complex behavioural hypotheses to be constructed and verified. As Hsiao (2007) adds, they provide greater control of the influence of omitted, unobservable variables. The last aspect, reported by Ashenfelter and Solon (1982), can be of great importance in particular when studying relationships between traits. This problem was considered, for example, in the work of Mellow (1981) and Mincer (1981). Panel studies also provide data to measure gross changes and aggregate individual data (Kowalczyk, 2001, p. 22). It should be noted, following Duncan and Kalton (1987), that it is possible to aggregate the sample over time with panel data, but only when the object of analysis is not a characteristic with static properties.

Sharot (1991), however, drew attention to the advantages that panel surveys have in terms of organisation and the process of conducting them themselves. Among these, The author mentioned the acquisition by the investigators of the ability to train respondents to perform complex tasks as part of the data collection process, such as completing a special diary. He also considered the possibility to collect more data than would be possible with a single interview to be a positive aspect and pointed out that longitudinal surveys allow significant costs to be spread over a longer period and a potentially large user base. Repeated contact between respondents and the interviewer and questionnaire also increases the chances that respondents will better understand the purpose of the survey, which may also translate into increased motivation to participate in the study and reliably complete the questionnaire (Duncan et al., 1986, p. 103). In some cases, the use of panel data can also simplify calculations and inference. Hsiao (2007) mentions as examples: non-stationary time series analysis (Anderson, 1959; Dickey and Fuller, 1979; 1981; Phillips and Durlauf, 1986), the problem of measurement error in the context of econometric model identification (Aigner et al., 1984; Biørn, 1992; Wansbeek and Koning, 1989) and the use of dynamic tobit models (Arellano et al., 1999).

According to Hsiao (1999), one of the advantages of longitudinal surveys is also that they provide a large number of observations, which allows for a reduction in collinearity between explanatory variables and an increase in the number of degrees of freedom, resulting in improved accuracy of estimates.

Surveys with complete rotation avoid some of the problems that arise with panel surveys. In this type of survey, it is possible to eliminate, first of all, the effect of survey fatigue and the suggestion of previous responses. The nature of this type of survey also avoids problems arising from changes in the composition of the population. In each subsequent survey round, the sample is drawn from the current population (Duncan and Kalton, 1987, p. 99). Kowalczyk (2001) points out that, as a result, these surveys can perform well in the estimation of population characteristics in particular periods and in the estimation of total values of characteristics for population, for all periods.

The complete replacement of sample elements in each survey also allows the sample to be cumulative by combining samples from different periods. This represents a major advantage for analyses involving rare events (Duncan and Kalton, 1987, p. 101).

2.4.2. Disadvantages and limitations of longitudinal surveys

A problem that may arise in the case of repeated surveys over time, which is of particular importance in the case of panel surveys, is the phenomenon called sample “attrition”. This phenomenon is associated with a reduction in the number of individuals from the original sample participating in the survey. One reason for attrition, according to Trivellato (1999), may be the refusal of an individual to participate in a subsequent survey. It therefore generates non-response at the level of individuals. The author mentions that non-random “attrition” of the sample, especially related to unobserved individual characteristics, may cause significant deviations in the analysed variables and lead to erroneous conclusions about the studied population. As reasons for the occurrence of non-response at the level of individuals for surveys conducted among households, Kowalczyk (2001) mentions illness, death and change of residence. When the units surveyed are businesses, these may include bankruptcy, merger and acquisition. The problem of the changing composition of the population is one of the main difficulties, or challenges, in conducting panel surveys. However, the problem of non-response can also arise at the level of individual questions (Trivellato, 1999, p. 342).

One of the considered in practice solutions to the problem of the changing composition of the population introduced in panel studies is the follow-up method. This method is used, among others, in the Social Diagnosis project, the Panel Study of Income Dynamics, and the German Socio-Economic Panel. In this method, when a person leaving a household in the sample creates a new household or joins an existing one, such households are included in the sample and all its members are surveyed. New members of households already in the survey are also added to the sample. The continuation method therefore takes into account both changes in the composition of households (changes in the relationships between household members), as well as their formation and division. Figure 2.10 shows an example of the changes in households that are taken into account by the above method (Ott, 1995, pp. 166–167). This solution allows, as Ott (1995) points out, an even broader analysis of the changes that occur in households over time, including, for example, the observation of the economic situation of both partners after divorce or of adult children leaving the family household.

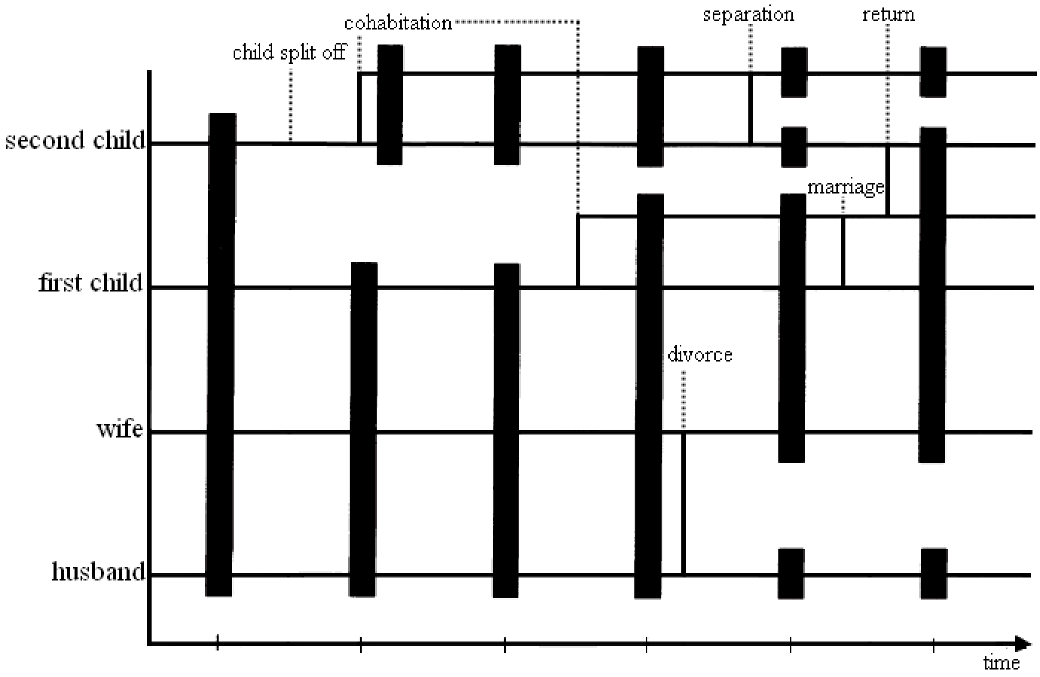


Figure 2.10. Scheme of the follow-up method

Source: Own elaboration based on Ott (1995, p. 167).

These issues were considered by Witte (1987, 1988), Giesecke (1989), and Witte and Lahmann (1988), among others. However, this method also has some disadvantages, it can lead, for example, to a significant increase in the sample size during a period when an intensive division of households takes place. Such a problem arose in the Panel Study of Income Dynamics conducted by the Institute for Social Research (University of Michigan). In this cases, it is necessary to take action to reduce the sample size. Its use also entails conducting the survey in a larger number of locations, which also entails increasing financial costs in subsequent waves. Procedures for 'tracking' panel survey units, which include the continuation method, are presented, among others, in the works of McAllister et al. (1973), Call et al. (1982), and Booth and Johnson (1985).

In practice, in order to counterbalance demographic inequalities arising during the "attrition" of the sample, as well as the effect of fatigue, a forced rotation of individuals is also often introduced (Sharot, 1991, p. 328). During rotation, respondents are discarded at a given maximum time of participation in the survey. The introduction of rotation, however, entails some negative consequences. These include, first and foremost, loss of sample continuity, which increases sampling errors, and the increased costs resulting from additional sampling, interviewing, recruitment, installation and removal of survey equipment. As Sharot (1991) adds, the associated increased burden on the survey contractor may distract from ensuring the quality of the survey being conducted.

In order to mitigate the effect of sample "attrition", a refreshing method is also used. The "refreshing" sample contains new randomly selected respondents who will receive the questionnaire at the same time as the second or subsequent waves of the panel. This method, as reported in Deng et al. (2013), is used by the National Center for Education Statistics in studies such as the Early Childhood Longitudinal Study and the National Educational Longitudinal Study, among others. The authors add that a similar refreshing effect is obtained by using partial rotation in panel studies.

Often, panel "attrition" is ignored and only data for those units that participated in all waves of the survey are used in the analyses. This approach assumes that panel "attrition" is missing completely at random (MCAR) and does not depend on either observed or unobserved variables. Another approach assumes that "attrition" is random but depends on observed variables (missing at random – MAR). One of the main methods in this approach is the reweighting method. It assumes that propensity to drop out is taken into account by assigning weights. Participation drop-out is therefore assumed to occur randomly in the weighting classes defined by the observed variables (Deng et al., 2013, pp. 239–240). An alternative approach to re-weighting is the single imputation method. In this method, each missing value is imputed according to a chosen procedure, e.g. hot deck or nearest neighbour. Thus, following Duncan and Kalton (1987), the weighting method is a global approach, whereas imputation treats each element individually. The last approach, as presented by Deng et al. (2013), is the assumption that, "attrition" of panels is not missing at random – NMAR) and depends on unobserved variables. In this case, the only way to obtain unbiased parameter estimators is to model missing data. However, model-based methods usually need to make strong assumptions about the "attrition" process, because most often the original sample does not contain enough information about the missing data mechanism.

Even when it is possible to obtain information from respondents, certain negative phenomena can occur that generate measurement errors and thus affect the quality of the data collected. One such phenomenon is the formation of so-called memory effects (memory errors). These are mainly due to the usually rather long period between successive rounds of surveys (Trivellato, 1990, p. 342). A phenomenon that may also affect the quality of the data obtained is the suggestion of previous answers or a kind of routine in answering questions, which may result, among other things, from survey fatigue. The length of time over which the survey is conducted can also have a major impact (Kowalczyk, 2001, p. 23).

Ott (1995) also draws attention to the problem of left and right censoring of data. In the first case, the event occurred or started before the study period, while in the second case it is known to occur or to end in the future but we do not know at what point. This problem occurs particularly with rare events.

A difficulty that can arise in panel studies as well as retrospective studies is the so-called "telescopic effect". This effect consists of remembering recent events as older – backward telescoping – or distant events as more recent – forward telescoping. In panel studies, however, according to Neter and Waksberg (1964), it is possible to reduce the effects of this through a method of tying events to the results of previous survey rounds, reminding subjects of their responses in previous interviews. In surveys with complete rotation, a telescoping effect can occur as in panel surveys, and in addition, data from this type of survey do not allow individual change to be measured or individual data to be aggregated (Duncan and Kalton, 1987, p. 101).

Touching on the distortions of information that may arise in the process of conducting longitudinal surveys, mention should be made of the initial interview. As Tanur (1981) points out, this interview is not used for comparison purposes in some panel studies. This action is to avoid the distortion of information that often occurs during the first interview. Sudman and Ferber (1970) provided examples indicating the reporting of higher levels of study characteristics at the beginning of the panel than in the second and subsequent periods.

Conducting studies that are repetitive over time also incurs greater financial costs. Duncan et al. (1986) point out that a survey consisting of two waves is already more costly than a cross-sectional survey extended with retrospective questions. The differences can reach 20 percent. Another important disadvantage pointed out by Sharot (1991) is the higher refusal rate in the first wave of the survey than in cross-sectional surveys. In the case of surveys with complete rotation, in addition to high costs, attention should be paid to their time-consuming nature, which is primarily associated with the need to draw samples in successive editions of the survey (Kowalczyk, 2001, p. 19).

2.5. Summary

This chapter discussed theoretical and practical aspects of single and longitudinal surveys. Subchapter 2.1 was focused on cross-sectional surveys. It discussed the essence of a statistical survey and the steps of its conduct. It also presented a classification of statistical surveys with some examples.

The next subsections covered the topics of longitudinal surveys. In subsection 2.2, the main reasons for the interest in longitudinal surveys and the development of this type of research were presented. The objectives of recurrent research over time were also discussed, along with their classification. The next subsection 2.3 addressed the classification of longitudinal surveys. The essence of balanced and unbalanced panel surveys, and partial and full rotation surveys was discussed in more detail. Surveys that are a combination of panel surveys and surveys with rotation were also presented. For each group, a scheme for carrying out this type of survey was presented. Selected examples of studies conducted both in Poland and worldwide were also described. Subchapter 2.4, however, dealt with the advantages and disadvantages of longitudinal surveys, as well as individual subgroups of repeated surveys over time identified in the previous subchapter. The benefits and limitations of conducting this type of research were also discussed. Selected methods to avoid or mitigate problems that may arise when conducting surveys over more than one period were also presented.

Chapter 3

Empirical best linear unbiased predictors

This chapter will discuss the problem of prediction using BLU- and EBLU-type predictors. The chapter will present the EBLU predictors proposed by Henderson (1950) and Royall (1976) in the light of the classification of linear mixed models into type A and B models. The author's proposal to use the EBLUP under the assumption of a linear mixed model with correlated random effects will also be presented. The chapter will also address the properties, possible modifications as well as applications of the presented predictors.

3.1. Empirical best linear unbiased predictors of Royall and Henderson

In this subsection, the empirical best linear unbiased predictors proposed by Royall (1976) and Henderson (1950) will be discussed. For both of the presented predictors, without loss of generality of consideration, we adopt the following decomposition of $\mathbf{Y}$:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_s^T & \mathbf{Y}_r^T \end{bmatrix}^T, \tag{3.1}$$

where $\mathbf{Y}_s$ and $\mathbf{Y}_r$ are vectors with n and $N-n$ elements for sampled and out-of-sample units, respectively. An analogous decomposition can be made of the matrix of auxiliary variables $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_s^T & \mathbf{X}_r^T \end{bmatrix}^T, \tag{3.2}$$

where the matrices $\mathbf{X}_s$ and $\mathbf{X}_r$ have dimensions $n \times p$ and $(N-n) \times p$, respectively, and the matrix $\mathbf{Z}$:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_s^T & \mathbf{Z}_r^T \end{bmatrix}^T, \tag{3.3}$$

where the matrices $\mathbf{Z}_s$ and $\mathbf{Z}_r$ have dimensions $n \times q$ and $(N-n) \times q$. The variance-covariance matrix $\mathbf{Y}$ can be written as:

$$\mathbf{V}(\boldsymbol{\delta}) = D^2(\mathbf{Y}) = D^2 \begin{bmatrix} \mathbf{Y}_s \\ \mathbf{Y}_r \end{bmatrix} = \begin{bmatrix} \mathbf{V}_{ss}(\boldsymbol{\delta}) & \mathbf{V}_{sr}(\boldsymbol{\delta}) \\ \mathbf{V}_{rs}(\boldsymbol{\delta}) & \mathbf{V}_{rr}(\boldsymbol{\delta}) \end{bmatrix}, \tag{3.4}$$

where $\boldsymbol{\delta}$ is a vector of unknown variational components. It should be added that the matrix $\mathbf{V}_{ss}(\boldsymbol{\delta})$ has dimensions $n \times n$, $\mathbf{V}_{rr}(\boldsymbol{\delta})$ – dimensions $(N-n) \times (N-n)$, and the dimensions of $\mathbf{V}_{sr}(\boldsymbol{\delta})$ and $\mathbf{V}_{rs}(\boldsymbol{\delta})$ are $n \times (N-n)$ and $(N-n) \times n$, respectively.

3.1.1. BLU predictor of Royall

When considering the BLUP proposed by Royall (1976), we make assumptions about the general linear model. In addition, we assume that the matrix $\mathbf{V}(\boldsymbol{\delta})$ is known. The problem under analysis is the prediction of a linear combination of the variable $\mathbf{Y}$, denoted as $\theta = \boldsymbol{\gamma}^T\mathbf{Y}$, which can also be written as follows:

$$\theta = \boldsymbol{\gamma}^T\mathbf{Y} = \boldsymbol{\gamma}^T\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}\right), \tag{3.5}$$

where the vector $\boldsymbol{\gamma}$ has the following form:

$$\boldsymbol{\gamma} = \begin{bmatrix}\boldsymbol{\gamma}_s^T & \boldsymbol{\gamma}_r^T\end{bmatrix}^T. \tag{3.6}$$

It should be added that when predicting the total value in the domain the k-th element of the vector, $\boldsymbol{\gamma}$ takes the value 1 for $k \in \Omega_d$ and 0 for $k \notin \Omega_d$. When the characteristic of interest is the mean value in the domain, the k-th element of the vector $\boldsymbol{\gamma}$ is N_d^{-1} for $k \in \Omega_d$ or 0 for $k \notin \Omega_d$.

According to Royall's (1976) theorem, the BLUP has the following form:

$$\hat{\theta}_{BLUP} = \boldsymbol{\gamma}_s^T\mathbf{Y}_s + \boldsymbol{\gamma}_r^T\left[\mathbf{X}_r\hat{\boldsymbol{\beta}}(\boldsymbol{\delta}) + \mathbf{V}_{rs}(\boldsymbol{\delta})\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\left(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}(\boldsymbol{\delta})\right)\right], \tag{3.7}$$

where:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\delta}) = \left(\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{X}_s\right)^{-1}\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{Y}_s. \tag{3.8}$$

This predictor is therefore ξ-unbiased and minimises the variance of the prediction error. If $\mathbf{R}_{sr} \neq \mathbf{0}$, that is, in particular, if we assume the diagonal form of the matrix $\mathbf{R}(\boldsymbol{\delta})$, following Żądło (2017), the predictor (3.7) simplifies to the following form:

$$\hat{\theta}_{BLUP} = \boldsymbol{\gamma}_s^T\mathbf{Y}_s + \boldsymbol{\gamma}_r^T\mathbf{X}_r\hat{\boldsymbol{\beta}}(\boldsymbol{\delta}) + \boldsymbol{\gamma}_r^T\mathbf{Z}_r\hat{\mathbf{v}}(\boldsymbol{\delta}), \tag{3.9}$$

where $\hat{\boldsymbol{\beta}}(\boldsymbol{\delta})$ and $\hat{\mathbf{v}}(\boldsymbol{\delta})$ are the vectors of fixed effects and random effects estimates, respectively, with $\hat{\boldsymbol{\beta}}(\boldsymbol{\delta})$ being given by the formula (3.8) and $\hat{\mathbf{v}}(\boldsymbol{\delta})$ being given by:

$$\hat{\mathbf{v}}(\boldsymbol{\delta}) = \mathbf{G}(\boldsymbol{\delta})\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\left(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}(\boldsymbol{\delta})\right). \tag{3.10}$$

Substituting in (3.7) for the vector $\boldsymbol{\delta}$ with its estimate, we obtain a two-stage predictor, called the empirical best linear unbiased predictor – EBLUP:

$$\hat{\theta}_{EBLUP} = \boldsymbol{\gamma}_s^T\mathbf{Y}_s + \boldsymbol{\gamma}_r^T\left[\mathbf{X}_r\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}) + \mathbf{V}_{rs}(\hat{\boldsymbol{\delta}})\mathbf{V}_{ss}^{-1}(\hat{\boldsymbol{\delta}})\left(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}})\right)\right], \tag{3.11}$$

where the vector $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}})$ is determined according to the formula (3.8), where $\boldsymbol{\delta}$ is replaced by $\hat{\boldsymbol{\delta}}$.

Under certain conditions, including those concerning the properties of the estimator of $\hat{\boldsymbol{\delta}}$ and the symmetry of the distributions of the effect vectors and random components, the predictor

(3.11) is ξ-unbiased. These conditions are presented more extensively in the work of Żądło (2004), together with a proof that is a generalisation of the proof of Kackar and Harwill (1981) for the predictor considered by Henderson (1950).

We can determine the mean squared error of a BLUP which has the form (3.7) based on the following formula (Royall, 1976):

$$MSE_{\xi}\left(\hat{\theta}_{BLUP}\right) = Var_{\xi}\left(\hat{\theta}_{BLU} - \theta\right) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}), \tag{3.12}$$

where

$$g_1(\boldsymbol{\delta}) = \boldsymbol{\gamma}_r^T\left(\mathbf{V}_{rr}(\boldsymbol{\delta}) - \mathbf{V}_{rs}(\boldsymbol{\delta})\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{V}_{sr}(\boldsymbol{\delta})\right)\boldsymbol{\gamma}_r \tag{3.13}$$

and

$$\begin{aligned} g_2(\boldsymbol{\delta}) = \boldsymbol{\gamma}_r^T\left(\mathbf{X}_r - \mathbf{V}_{rs}(\boldsymbol{\delta})\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{X}_s\right)\left(\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{X}_s\right) \times \\ \times\left(\mathbf{X}_r - \mathbf{V}_{rs}(\boldsymbol{\delta})\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{X}_s\right)^T\boldsymbol{\gamma}_r^T. \end{aligned} \tag{3.14}$$

For the empirical version of Royall's (1976) predictor, given by the formula (3.11), the MSE, with some additional assumptions presented by Żądło (2009), has the following form:

$$MSE_{\xi}\left(\hat{\theta}_{EBLUP}\right) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) + g_3(\boldsymbol{\delta}) + o(D^{-1}), \tag{3.15}$$

where $g_1(\boldsymbol{\delta})$ and $g_2(\boldsymbol{\delta})$ are determined according to the formulas (3.13) and (3.14), respectively. The last component of $g_3(\boldsymbol{\delta})$, if the relevant assumptions are met, has the form (Żądło, 2009, p. 109):

$$g_3(\boldsymbol{\delta}) = \operatorname{tr}\left(\frac{\partial\mathbf{c}^T}{\partial\boldsymbol{\delta}}\mathbf{V}_{ss}(\boldsymbol{\delta})\frac{\partial\mathbf{c}^T}{\partial\boldsymbol{\delta}}\breve{\mathbf{D}}^2(\hat{\boldsymbol{\delta}})\right), \tag{3.16}$$

where $\mathbf{c}^T = \boldsymbol{\gamma}_r^T\mathbf{V}_{rs}(\boldsymbol{\delta})\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})$, and $\breve{\mathbf{D}}^2(\hat{\boldsymbol{\delta}})$ is the asymptotic variance-covariance matrix of the estimator of $\hat{\boldsymbol{\delta}}$. The estimation problem of $MSE_{\xi}\left(\hat{\theta}_{EBLUP}\right)$ will be discussed in the third subsection of this chapter.

Royall's (1976) predictor was considered in the work of Chandra and Chambers (2006), among others, in the context of economic aspects of agriculture. The authors considered a dataset presented in the work of Chandra and Chambers (2005). A population of farms of $N = 81,982$ was analysed, from which a sample size of $n = 1,652$ was drawn. The farms were further divided into 29 domains defined by agricultural regions. Eight variables were included in the analyses, including total farm operating costs, area under crops (in hectares), number of cattle on the farm, number of sheep on the farm, total equity, and total farm debt. The mean value in the domain of each variable was estimated. The area of the farm was used as an auxiliary variable. The simulation studies carried out compared the properties of the EBLUP and the direct predictor. It should be added that the mean squared error estimator for the EBLUP and the direct predictor were determined based on the estimator proposed by Prasad and Rao (1990) and used the robust method presented in the work of Chambers and Chandra (2006). For the *rRMSE* predictors of the mean across domains and their median, the results obtained for the EBLUP were better or close to those of the direct predictor.

3.1.2. Henderson's best linear unbiased predictor

For the predictor proposed by Henderson (1950), we make assumptions about the general linear mixed model given by the formula (1.89). In this case we consider the prediction of a linear combination of $\boldsymbol{\beta}$ and $\mathbf{v}$, which can be written as $\theta^s = \mathbf{l}^T \boldsymbol{\beta}(\boldsymbol{\delta}) + \mathbf{m}^T \mathbf{v}(\boldsymbol{\delta})$, where $\mathbf{l}^T$ and $\mathbf{m}^T$ are known. When the aim is to predict a linear combination of the variable Y, then $\mathbf{l}^T = \boldsymbol{\gamma}^T \mathbf{X}$ and $\mathbf{m}^T = \boldsymbol{\gamma}^T \mathbf{Z}$ are assumed. It should be noted, however, that in this case, we do not predict $\theta = \boldsymbol{\gamma}^T \mathbf{Y}$, as is the case when using the predictor proposed by Royall (1976).

According to Henderson's (1950) theorem, the best linear unbiased predictor is given by the formula:

$$\hat{\theta}^s_{BLUP} = \mathbf{l}^T \hat{\boldsymbol{\beta}}(\boldsymbol{\delta}) + \mathbf{m}^T \hat{\mathbf{v}}(\boldsymbol{\delta}), \tag{3.17}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{v}}$ are given by the formulae (3.8) and (3.10). It should be added that the predictor (3.17) is a special case of the predictor proposed by Royall (1976) because:

- Royall (1976) carries out his considerations under the assumption of a general linear model, which is a generalisation of the general linear mixed model considered by Henderson (1950),
- Royall (1976) considers the prediction $\theta = \boldsymbol{\gamma}^T \mathbf{Y}$, which is a generalisation of the characteristic considered by Henderson (1950) θ_s, as:

$$\theta = \boldsymbol{\gamma}^T \mathbf{Y} = \boldsymbol{\gamma}^T \mathbf{X} \boldsymbol{\beta}(\boldsymbol{\delta}) + \boldsymbol{\gamma}^T \mathbf{Z} \mathbf{v}(\boldsymbol{\delta}) + \boldsymbol{\gamma}^T \mathbf{e} = \theta^s + \boldsymbol{\gamma}^T \mathbf{e}, \tag{3.18}$$

where $\boldsymbol{\gamma}^T \mathbf{X} = \mathbf{l}^T$ and $\boldsymbol{\gamma}^T \mathbf{Z} = \mathbf{m}^T$.

As in the case of Royall's (1976) predictor, by replacing the vector $\boldsymbol{\delta}$ in (3.17) with its estimate, we can obtain an empirical version of this predictor:

$$\hat{\theta}_{EBLUP} = \mathbf{l}^T \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}) + \mathbf{m}^T \hat{\mathbf{v}}(\hat{\boldsymbol{\delta}}), \tag{3.19}$$

where $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}})$ and $\hat{\mathbf{v}}(\hat{\boldsymbol{\delta}})$ are determined based on the formulae (3.8) and (3.10) for $\hat{\boldsymbol{\delta}}$. According to the theorem presented in the paper by Kackar and Harville (1981), the predictor given by the formula (3.19) is ξ-unbiased when the assumptions of the general mixed model are satisfied and when $E_\xi\left(\hat{\theta}_{EBLUP}\right)$ is finite, $\hat{\boldsymbol{\delta}}$ is an estimator having the evenness property $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s) = \hat{\boldsymbol{\delta}}(-\mathbf{Y}_s)$ and invariability against displacement: $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s - \mathbf{X}_s \mathbf{b}) = \hat{\boldsymbol{\delta}}(\mathbf{Y}_s)$, and the distributions of the effects and random components are symmetric with respect to zero. The above properties of the $\hat{\boldsymbol{\delta}}$

estimator are ensured, among other things, by the use of the maximum likelihood method and the restricted maximum likelihood method.

The mean squared error of the predictor proposed by Henderson (1950) is given by the formula (cf. Rao and Molina, 2015, p. 101):

$$MSE\left(\hat{\theta}^{s}_{BLUP}\right) = g^{s}_{1}(\delta) + g^{s}_{2}(\delta), \tag{3.20}$$

where

$$g^{s}_{1}(\boldsymbol{\delta}) = \mathbf{m}^{T}\left(\mathbf{G}(\boldsymbol{\delta}) - \mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^{T}_{s}\mathbf{V}^{-1}_{ss}(\boldsymbol{\delta})\mathbf{Z}_{s}\mathbf{G}(\boldsymbol{\delta})\right)\mathbf{m} \tag{3.21}$$

and

$$\begin{aligned} g^{s}_{2}(\boldsymbol{\delta}) = {} & \left(\mathbf{l} - \mathbf{m}^{T}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^{T}_{s}\mathbf{V}^{-1}_{ss}(\boldsymbol{\delta})\mathbf{X}_{s}\right)\left(\mathbf{X}^{T}_{s}\mathbf{V}^{-1}_{ss}(\boldsymbol{\delta})\mathbf{X}_{s}\right)^{-1} \times \\ & \times \left(\mathbf{l} - \mathbf{m}^{T}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^{T}_{s}\mathbf{V}^{-1}_{ss}(\boldsymbol{\delta})\mathbf{X}_{s}\right)^{T}. \end{aligned} \tag{3.22}$$

The mean squared error of the empirical version of Henderson's (1950) predictor, with some additional assumptions presented in Datta and Lahiri (2000), is given by the formula:

$$MSE_{\xi}\left(\hat{\theta}^{s}_{EBLUP}\right) = g^{s}_{1}(\boldsymbol{\delta}) + g^{s}_{2}(\boldsymbol{\delta}) + g^{s}_{3}(\boldsymbol{\delta}) + o(D^{-1}), \tag{3.23}$$

where $g^{s}_{1}(\boldsymbol{\delta})$ and $g^{s}_{2}(\boldsymbol{\delta})$ are determined from (3.21) and (3.22), respectively. The value of the last component, as shown by Datta and Lahiri (2000), can be determined by the following formula:

$$g^{s}_{3}(\boldsymbol{\delta}) = \mathrm{tr}\left(\frac{\partial \mathbf{b}^{T}}{\partial \boldsymbol{\delta}}\mathbf{V}_{ss}(\boldsymbol{\delta})\frac{\partial \mathbf{b}^{T}}{\partial \boldsymbol{\delta}}\breve{\mathbf{D}}^{2}(\hat{\boldsymbol{\delta}})\right), \tag{3.24}$$

where $\mathbf{b}^{T} = \mathbf{m}^{T}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^{T}_{s}\mathbf{V}^{-1}_{ss}(\boldsymbol{\delta})$ and $\breve{\mathbf{D}}^{2}\left(\hat{\boldsymbol{\delta}}\right)$ is the asymptotic variance-covariance matrix $\hat{\boldsymbol{\delta}}$. The issue of the estimation of $MSE_{\xi}\left(\hat{\theta}^{s}_{EBLUP}\right)$ is dedicated to the third subsection of this chapter.

The application of the EBLUP proposed by Henderson (1950) in the context of analyses of an economic nature, more specifically analyses of firms, is presented in the paper by Ghosh and Rao (1994), among others. The authors addressed the problem of prediction using the EBLUP of corporate wages with gross business income as an auxiliary variable. They conducted their considerations based on an artificial population designed to resemble the dataset in the work of Särndal and Hidiroglou (1989). The authors considered dividing the population into 16 domains. The population and sample sizes considered were $N = 144$ and $n = 38$, respectively. It should be added that the analyses used a model with a nested random component. Their analyses compared the properties of the EBLUP with, among others, a hierarchical Bayesian predictor and a synthetic ratio estimator. Ghosh and Rao (1994) used the mean squared error, the relative root of the mean squared error and the mean prediction error, among others. The authors noted that the results for the EBLUP and the hierarchical Bayesian predictor were similar. Furthermore, significantly better results for mean squared error were obtained for these predictors than for the other predictors and estimators considered.

3.2. Empirical best linear unbiased predictor and classification of linear mixed models

In this subsection, the EBLUP will be discussed in the context of the division of linear mixed models into type A and type B models, which were discussed more extensively in subsection 1.2. For the former group of models, the EBLUP assuming the Fay–Herriot model (Fay–Herriot, 1979) is presented, and for the latter, the model considered by Battese et al. (1988). The subsection also presents the form of the mean squared error for the discussed predictors. The considerations are carried out on models assumed for cross-sectional data.

3.2.1. EBLUP for type A models

Considering the classification of linear mixed models, the EBLUP can be divided into those based on models belonging to type A as well as models of type B. For the first class of models, one of the most commonly used is the model proposed by Fay and Herriot (1979) given by the formula (1.92). This model is discussed further in subsection 1.2.

The EBLUP of the domain mean value assuming the Fay–Herriot model (FH-EBLUP), in which the normality of the distribution of both components and random effects is assumed, can be written as (cf. Rao and Molina, 2015, p. 126):

$$\hat{\theta}_d^{FH}\left(\hat{\sigma}_v^2\right) = \mathbf{x}_d^T\hat{\boldsymbol{\beta}} + \hat{\gamma}_d\left(\hat{\theta}_d - \mathbf{x}_d^T\hat{\boldsymbol{\beta}}\right) = \hat{\gamma}_d\hat{\theta}_d + (1-\hat{\gamma}_d)\,\mathbf{x}_d^T\hat{\boldsymbol{\beta}}, \tag{3.25}$$

where:

$$\hat{\gamma}_d = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2}, \tag{3.26}$$

and $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$ obtained using a weighted least squares method with weights $\left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2\right)$. This predictor is therefore the sum of $\hat{\theta}_d$ – the direct θ estimator and the synthetic model-based regression estimator. Note that the FH-EBLUP gives more weight to the synthetic component $\mathbf{x}_d^T\hat{\boldsymbol{\beta}}$ for high values of the variance of $\hat{\sigma}_e^2$ or low values of the variance of random effects. This predictor, like the Fay–Herriot model itself, allows data from different sources to be combined. For domains not in the sample, the predictor is reduced to the synthetic regression estimator $\mathbf{x}_d^T\hat{\boldsymbol{\beta}}$. The information on the additional feature values for these domains is used for this purpose. This makes it possible, among other things, to include geographical variables – the coordinates of the centres of gravity – in the analysis.

The FH-EBLUP is effective under the assumption of normality in a linear mixed model. However, it is possible to extend this predictor to cases where the dependence is not linear, as was considered, among others, in the work of Giusti et al. (2012). It is also possible to include correlations between random effects in this predictor, as was presented in the works of Petrucci and Salvati (2006), and Pratesi and Salvati (2008, 2009). This predictor is also consistent after the sampling design as Pratesi (2015) points out, however, it is not robust to the presence of outliers.

The mean squared error of the above predictor is given by the formula (cf. Prasad and Rao, 1990, p. 165):

$$MSE\left(\hat{\gamma}_d\right) = g_{1d}\left(\sigma_v^2\right) + g_{2d}\left(\sigma_v^2\right) + g_{3d}\left(\sigma_v^2\right), \tag{3.27}$$

where $g_{1d}\left(\sigma_v^2\right) = \gamma_d \sigma_v^2$, $g_{2d}\left(\sigma_v^2\right) = \left(1-\gamma_d\right)^2 \mathbf{x}_d^T Var\left(\hat{\boldsymbol{\beta}}\right)\mathbf{x}_d$, and $g_{3d}\left(\sigma_v^2\right) = \frac{\sigma_e^4}{\left(\hat{\sigma}_v^2+\sigma_e^2\right)^3} Var\left(\sigma_v^2\right)$. The estimator proposed by Prasad and Rao (1990), estimators using the jackknife method, or the parametric bootstrap method, discussed in more detail in the next subsection of this monograph, can be used to estimate it.

The FH-EBLUP has found application in economic aspects of agriculture, among other things. The work of Sud et al. (2011) used this predictor to estimate average rice yields at the district level in Uttar Pradesh, India. The study used data from 2009/2010 for 58 districts. Information on additional population variables came from the 2001 census. These variables were average household size and female population.

3.2.2. EBLUP for type B models

One of the most commonly used models belonging to type B in prediction using BLUP is that considered by Battese et al. (1988). This model is described in more detail in subsection 1.2 on the classification of overpopulation models. The BLUP of the domain mean value based on this model is given by the following formula (Rao and Molina, 2015, pp. 174–177):

$$\theta_d^{BHF}\left(\sigma_v^2\right) = \hat{\gamma}_d\left[\bar{y}_{da} + \left(\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da}\right)^T \tilde{\boldsymbol{\beta}}\right] + \left(1-\hat{\gamma}_d\right)\bar{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}}, \tag{3.28}$$

where

$$\hat{\gamma}_d = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2/a_{d\cdot}}, \qquad \bar{y}_{da} = \sum_{i=1}^{n_d} \frac{a_{di} y_{di}}{a_{d\cdot}}, \qquad \bar{\mathbf{x}}_{da} = \sum_{i=1}^{n_d} \frac{a_{di}\mathbf{x}_{di}}{a_{d\cdot}}, \qquad a_{d\cdot} = \sum_{i=1}^{n_d} a_{di},$$

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \mathbf{X}_d^T \mathbf{V}_d^{-1}\left(\boldsymbol{\delta}\right)\mathbf{X}_d\right)^{-1}\left(\sum_{d=1}^{D} \mathbf{X}_d^T \mathbf{V}_d^{-1}\left(\boldsymbol{\delta}\right)\mathbf{Y}_d\right),$$

and $\mathbf{X}_d$ is a vector of known mean values. It should be added that the weights $\hat{\gamma}_d$ take values in the interval $(0,1)$ and measure the relative variation between domains, and the values of a_{di} are determined as $a_{di} = k_{di}^2$, where k_{di} are weights to allow for heteroscedasticity of the random components. An empirical version of the predictor given by the formula (3.28) is obtained by replacing σ_v^2 and σ_e^2 with their respective assessments.

The mean squared error of the above predictor is determined based on the following formula (Rao and Molina, 2015, pp. 175–176):

$$MSE\left(\theta_d^{BHF}\left(\sigma_v^2\right)\right) = g_1^B(\sigma_v^2) + g_2^B(\sigma_v^2), \tag{3.29}$$

where:

$$g_1^B(\sigma_v^2) = \gamma_d\left(\sigma_e^2/a_{d\cdot}\right), \tag{3.30}$$

$$g_2^B(\sigma_v^2) = \left(\bar{\mathbf{X}}_d - \gamma_d\bar{\mathbf{x}}_{da}\right)^T\left(\sum_{d=1}^{D}\mathbf{A}_d\right)^{-1}\left(\bar{\mathbf{X}}_d - \gamma_d\bar{\mathbf{x}}_{da}\right), \tag{3.31}$$

and:

$$\mathbf{A}_d = \mathbf{X}_d^T\mathbf{V}_d^T\mathbf{X}_d = \sigma_e^{-2}\left(\sum_{i=1}^{n_d} a_{di}\mathbf{x}_{di}\mathbf{x}_{di}^T - \gamma_d\sum_{i=1}^{n_d} a_{di}\bar{\mathbf{x}}_{da}\bar{\mathbf{x}}_{da}^T\right). \tag{3.32}$$

A predictor of the BHF-EBLUP type is considered in the paper by Prasad and Rao (1990), among others, which addresses a problem in the economic aspects of agriculture. In a simulation study, the authors considered data generated from a maize crop dataset from the work of Battese et al. (1988). The paper considers not only the problem of prediction using the EBLUP but also the issue of estimating the mean squared error of the prediction. This issue will be discussed in more detail in the next subsection of this paper.

3.3. EBLUP and the class of linear mixed models with correlated random effects

It should be added that it is also possible to apply the EBLUP having the form (3.11) in the case of a linear mixed model with correlated random effects vectors given by the formula (1.131). For the above proposal, this predictor has the form:

$$\hat{\theta}_{EBLUP}^{*} = \boldsymbol{\gamma}_s^T\mathbf{Y}_s + \boldsymbol{\gamma}_r^T\left[\mathbf{X}_r\hat{\boldsymbol{\beta}}^*(\hat{\boldsymbol{\delta}}) + \mathbf{V}_{rs}^*(\hat{\boldsymbol{\delta}})\mathbf{V}_{ss}^{*\,-1}(\hat{\boldsymbol{\delta}})\left(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}^*(\hat{\boldsymbol{\delta}})\right)\right], \tag{3.33}$$

where $\mathbf{V}_{ss}^*(\hat{\boldsymbol{\delta}})$, $\mathbf{V}_{rs}^*(\hat{\boldsymbol{\delta}})$ and $\hat{\boldsymbol{\beta}}^*(\hat{\boldsymbol{\delta}})$ are determined according to the model (3.11), where the variance-covariance matrix $\mathbf{Y}$ is given by the formula (1.133).

The above problem was considered in the work of Krzciuk (2020). The analyses studied the properties of the EBLUP under the assumption of a model with correlated and uncorrelated random effects vectors in the case of correct and incorrect model specification. Simulation studies were conducted based on data presented in Särndal et al. (1992) about Swedish counties. The results indicate good properties of the proposed EBLUP, assuming a model with correlated random effects vectors. The simulation-derived relative bias values for this predictor did not exceed 1%. When the data were generated according to the model with correlated random effects vectors, the increase in accuracy for individual domains was between 4% and 39% compared to the non-correlated predictor. In the second part of the simulation study, when no correlation between the random effects vectors was taken into account in the data generation process, the loss of accuracy for the proposed predictor was no more than 5% relative to the case with the correct model specification. The analyses also considered the problem of estimating the mean squared error. The use of estimators based on the parametric bootstrap method was proposed.

3.4. Mean squared prediction error and its estimator

When considering the problem of predicting characteristics in the domain using the EBLUP, it is also necessary to address the issue of estimating the mean squared error of this class of predictors. Among the classical estimators of the mean squared error of the EBLUP, we can mention the naive estimator, which is given by the following formula (Kackar and Harville, 1984, pp. 854–855):

$$\widehat{MSE}_N(\hat{\theta}^{EBLUP}) = g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}). \tag{3.34}$$

It is therefore in the form of the BLUP mean squared error (cf. formulas (3.12) and (3.20)), where the vector $\boldsymbol{\delta}$ is replaced by its estimator. In this group, we can also distinguish the estimator proposed by Datta and Lahiri (2000). Considering the case where the model parameters were estimated using the REML method, this estimator has the following form:

$$\widehat{MSE}_{DL}(\hat{\theta}^{EBLUP}) = g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) + 2g_3(\hat{\boldsymbol{\delta}}), \tag{3.35}$$

where $g_1(\hat{\boldsymbol{\delta}})$, $g_2(\hat{\boldsymbol{\delta}})$ and $g_3(\hat{\boldsymbol{\delta}})$ are determined from the formulae (3.13), (3.14) and (3.16) for the predictor of Royall (1976) and (3.21), (3.22) and (3.24) for the predictor of Henderson (1950). Also in this case, the vector $\boldsymbol{\delta}$ is replaced by its estimator. It should be added that the bias of the naive estimator is of order $O(D^{-1})$, while that of the estimator of Datta and Lahiri (2000) is of order $o(D^{-1})$. It should be added, following Fuller (1976), that for sequences of positive numbers $\{r_n\}$ and $\{c_n\}$, $c_n = O(r_n)$ means that c_n is of order r_n if there exists a positive real number M such that $|c_n| \leqslant Mr_n$ for every n and $c_n = o(r_n)$ means that c_n is of an order lower than r_n when $\lim_{n\to\infty} \frac{c_n}{r_n} = 0$

Among the estimators of the mean squared error of the EBLUP, there are also estimators using the jackknife method. The estimator proposed by Jiang et al. (2002) is given by the following formula:

$$\widehat{MSE}_{JACK1}(\hat{\theta}^{EBLUP}) = g_1(\hat{\boldsymbol{\delta}}) - \frac{D-1}{D}\sum_{d=1}^{D}\Big(g_1(\hat{\boldsymbol{\delta}}_{-d}) - g_1(\hat{\boldsymbol{\delta}})\Big) + \\ + \sum_{d=1}^{D}\Big(\hat{\theta}_{EBLUP}(\hat{\boldsymbol{\delta}}_{-d}) - \hat{\theta}_{EBLUP}(\hat{\boldsymbol{\delta}})\Big)^2, \tag{3.36}$$

where $\hat{\boldsymbol{\delta}}_{-d}$ is estimated based on the out-of-sample data excluding the d-th domain, and $g_1(\hat{\boldsymbol{\delta}})$ is calculated analogously to the previously presented estimators. It should be added that this estimator is asymptotically unbiased, and the order of its bias is equal to $o(D^{-1-\varepsilon})$ $(0 < \varepsilon < 0.5)$.

Chen and Lahiri (2002, 2003) proposed a modification of the estimator of Jiang et al. (2002) due to the potential for negative values of this estimator. The estimator considered by the authors can be written as:

$$\widehat{MSE}_{JACK2}(\hat{\theta}^{EBLUP}) = g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) + \sum_{d=1}^{D} w_d \left(\hat{\theta}_{EBLUP}(\hat{\boldsymbol{\delta}}_{-d}) - \hat{\theta}_{EBLUP}(\hat{\boldsymbol{\delta}}) \right)^2 - \\ - \frac{D-1}{D} \sum_{d=1}^{D} w_d \left(g_1(\hat{\boldsymbol{\delta}}_{-d}) + g_2(\hat{\boldsymbol{\delta}}_{-d}) - g_1(\hat{\boldsymbol{\delta}}) - g_2(\hat{\boldsymbol{\delta}}) \right). \quad (3.37)$$

Note that the choice of weights w_d is not straightforward. Chen and Lahiri (2003) suggest weights for the Fay–Herriot model of the form $w_d = \frac{D-1}{D}$ or $w_d = \mathbf{x}_d^T \left(\sum_{u=1}^{D} \mathbf{x}_u \mathbf{x}_u^T \right) \mathbf{x}_d$. The bias of this estimator, as given by Chen and Lahiri (2003), is of the order of $O(D^{-1})$.

Another group of estimators presented in this paper are those based on the parametric bootstrap method. These estimators are based on the following bootstrap model (cf. Chatterjee et al., 2008, pp. 1229–1230):

$$\mathbf{Y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{v}^* + \mathbf{e}^*, \quad (3.38)$$

where: $\mathbf{v}^* \sim N(\mathbf{0}, \mathbf{G}(\hat{\boldsymbol{\delta}}))$, $\mathbf{e}^* \sim N(\mathbf{0}, \mathbf{R}(\hat{\boldsymbol{\delta}}))$; $\hat{\boldsymbol{\delta}}$ is the $\boldsymbol{\delta}$ estimator obtained by the REML or ML method, and $\hat{\boldsymbol{\beta}}$ is the $\boldsymbol{\beta}$ estimator obtained by the least squares method. Among the estimators using the parametric bootstrap method, we can distinguish, e.g., the estimator considered in the work of González-Manteiga et al. (2008):

$$M\hat{S}E^{boot}(\hat{\theta}^{EBLUP}) = E_*(\hat{\theta}^{EBLUP}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}^*), \hat{\boldsymbol{\delta}}^*) - \theta^*)^2 \\ = B^{-1} \sum_{b=1}^{B} (\hat{\theta}^{EBLUP}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}^{*(b)}), \hat{\boldsymbol{\delta}}^{*(b)}) - \theta^{*(b)})^2, \quad (3.39)$$

where $\hat{\boldsymbol{\delta}}^{*(b)}$ is given by the same formula as $\boldsymbol{\delta}$, where $\mathbf{Y}$ is replaced by $\mathbf{Y}^*$, $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\beta}}$ are estimators obtained by the REML method, $\theta^{*(b)}$ is the value of the characteristic of interest θ obtained in the b-th implementation of the model (3.38), $E^*(.)$ is the expected value in the bootstrap distribution. The second of the MSE estimators presented in this paper based on the parametric bootstrap method is the estimator proposed in Butar and Lahiri (2003):

$$\widehat{MSE}^{boot-BL}(\hat{\theta}^{EBLUP}) = g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) + E_*(g_1(\hat{\boldsymbol{\delta}}^*) + g_2(\hat{\boldsymbol{\delta}}^*) - (g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}))) + \\ + E_*(\hat{\theta}^{EBLUP}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}^*), \hat{\boldsymbol{\delta}}^*) - \hat{\theta}^{EBLUP}(\hat{\boldsymbol{\delta}}))^2, \quad (3.40)$$

where $g_1(\hat{\boldsymbol{\delta}}^*)$ and $g_2(\hat{\boldsymbol{\delta}}^*)$ are determined analogously to the other estimators. It should be added that this estimator is asymptotically unbiased:

$$E_{\xi}(\widehat{MSE}^{boot-BL}(\hat{\theta}^{EBLUP})) - MSE_{\xi}(\hat{\theta}^{EBLUP}) = \mathrm{o}(D^{-1}). \tag{3.41}$$

It is also possible to use mean squared error estimators using a non-parametric bootstrap method. This problem was considered by Pfeffermann and Glickman (2004), among others. Hall and Maiti (2006) proposed the use of a double parametric bootstrap method for estimating the mean squared error (see Algorithm 1). Other MSE estimators derived by this method and allowing for bias correction are also considered in the literature, e.g. Pfeffermann and Correa (2012).

It should be added that the mean squared error estimators discussed in this subsection can also be used to assessment the MSE of the EBLUP, taking into account the correlation between the random effects vectors, given by the formula (3.33).

Algorithm 1. Double parametric bootstrap method

1. Generate B_1 new populations and samples according to the assumed model and determine the value of the considered predictor. The MSE estimator at the first step of the procedure has the form (Hall and Maiti, 2006, p. 226):

$$\widehat{MSE}_1^{dboot}(\hat{\theta}^{EBLUP}) = B_1^{-1}\sum_{b_1=1}^{B_1}\left(\hat{\theta}_{b_1}^{EBLUP} - \hat{\theta}_{b_1}\right)^2. \tag{3.42}$$

2. For the samples generated in step one, the calculations performed in step one are repeated B_2 times. The MSE estimator from the second step is therefore of the form (Hall and Maiti, 2006, p. 226):

$$\widehat{MSE}_2^{dboot}(\hat{\theta}^{EBLUP}) = B_1^{-1}\sum_{b_1=1}^{B_1} B_2^{-1}\sum_{b_2=1}^{B_2}\left(\hat{\theta}_{b_2}^{EBLUP} - \hat{\theta}_{b_2}\right)^2. \tag{3.43}$$

3. Calculation of the MSE estimator using the above bootstrap method procedure:

$$\widehat{MSE}^{dboot}(\hat{\theta}^{EBLUP}) = 2\widehat{MSE}_1^{dboot}(\hat{\theta}^{EBLUP}) - \widehat{MSE}_2^{dboot}(\hat{\theta}^{EBLUP}). \tag{3.44}$$

However, it is necessary to consider the form of the variance-covariance matrix for the aforementioned model, which has the form according to the formula (1.133). This necessitates the estimation of the correlation coefficient ρ.

3.5. Selected EBLUP modifications and their applications

In addressing the issue of prediction using EBLUPs, it is important to mention the many modifications of these predictors. Among the selected modifications presented in this subsection are the EBLU predictor taking into account the spatial correlation of random effects, the geographically weighted EBLUP, the pseudo-EBLUP, and the non-parametric EBLUP. Robust versions of some of the above predictors are also presented. The description of each of the presented predictors is also supplemented with issues concerning the mean squared error of the prediction.

3.5.1. SEBLUP

A modification of the empirical best linear unbiased predictor denoted as SEBLUP (Spatial empirical best linear unbiased predictor) allows prediction when random effects correlations are present. This predictor is presented, among others, in the work of Molina et al. (2009) and Singh et al. (2005). This generalisation can be applied to type A as well as type B models. Following Molina et al. (2009), a Simultaneous Spatial Autoregressive Process (SAR process) is incorporated into the models.

For type A models, which include the Fay–Herriot model given by the formula (1.92), the correlated random effects vector $\mathbf{v}$ is assumed to be of the form (Cressie, 1993):

$$\mathbf{v} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u}, \tag{3.45}$$

where $\mathbf{u}$ is a D-element vector of independent random effects with variance σ_u^2, and ρ is an unknown parameter. The matrix of spatial weights $\mathbf{W}$ is of dimension $D \times D$, since correlation is assumed between domains rather than between population elements. It should be noted that the proximity of domains can be considered not only in a geographical sense but also in an economic sense. In a geographical sense, the creation of a weight matrix can take into account, among other things, whether objects share a common boundary (Karpuk, 2015) or the length of a common boundary (Dacey, 1968). In an economic sense, we can use variables such as the unemployment rate or the value of investments (Pietrzak, 2010). The $\mathbf{W}$ weight matrix can also be based on the values of mutual trade transactions, capital flows and migration between spatial units (Conley, 1999). It should be added that the rows of the $\mathbf{W}$ matrix are usually standardised. The problem of defining the weights matrix has been extensively presented in a book edited by Suchecki (2010). The variance-covariance matrix of random effects $\mathbf{G}$ is given in this case by the formula (Molina et al., 2008, p. 444):

$$\mathbf{G} = \sigma_u^2\left[(\mathbf{I} - \rho\mathbf{W})\left(\mathbf{I} - \rho\mathbf{W}^T\right)\right]^{-1}. \tag{3.46}$$

The $\mathbf{R}$ matrix can be written as:

$$\mathbf{R} = \sigma_e^2\mathbf{I}. \tag{3.47}$$

By substituting (3.46) and (3.47) into (1.89), we obtain a matrix $\mathbf{V}$ of the form (Pratesi and Salvati, 2008, p. 116):

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R} = \mathbf{Z}\sigma_u^2 \left[(\mathbf{I} - \rho\mathbf{W})\left(\mathbf{I} - \rho\mathbf{W}^T\right)\right]^{-1} \mathbf{Z}^T + \sigma_e^2\mathbf{I}. \tag{3.48}$$

Taking into account the (3.45) in (1.92), the model used for prediction using the SEBLUP can be written as (Pratesi and Salvati, 2008, p. 115):

$$\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u} + \mathbf{e}, \tag{3.49}$$

where $\hat{\boldsymbol{\theta}}$ is the vector of direct estimators of the characteristic vector $\boldsymbol{\theta}$, $\mathbf{X}$ and $\mathbf{Z}$ are the matrices of the auxiliary variables, $\boldsymbol{\beta}$ is the vector p of unknown parameters, and $\mathbf{e}$ is the vector of random components.

The predictor of type Spatial EBLU for the model (3.49) can therefore be written as (Pratesi and Salvati, 2008, p. 116):

$$\tilde{\theta}_d^{FH-SEBLUP}\left(\hat{\sigma}_u^2, \hat{\rho}\right) = \mathbf{x}_d\hat{\boldsymbol{\beta}} + \mathbf{b}_d^T\hat{\mathbf{G}}\mathbf{Z}^T\left\{\mathbf{R} + \mathbf{Z}\left(\hat{\mathbf{G}}\mathbf{Z}^T\right)^{-1}\left(\hat{\boldsymbol{\theta}} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)\right\}, \tag{3.50}$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\hat{\boldsymbol{\theta}}$, $\mathbf{b}_d^T$ is a D-element vector containing $D-1$ zeros and 1 as the d-th element. In addition, the matrix $\mathbf{R}$ is given by the formula (3.47) and $\hat{\mathbf{G}} = \mathbf{G}\left(\hat{\sigma}_u^2, \hat{\rho}\right)$, where $\mathbf{G}$ is determined according to the formula (3.46). Following Pratesi and Salvati (2008), it should be added that, under the assumption of normality of random effects, the parameters σ_u^2 and ρ can be estimated using either the maximum likelihood method (ML) or the restricted maximum likelihood method (REML). However, only the REML method takes into account the loss of degrees of freedom resulting from the estimation of $\boldsymbol{\beta}$. The problem of estimating random effects variance is addressed, among others, in Siswantining et al. (2020).

The SEBLUP for models belonging to type A are effective under the assumption of normality and spatial correlation in a linear mixed model. The advantage of this predictor, as with the EBLUP, is furthermore that it does not require access to microdata at the individual level. It does, however, allow for the inclusion of random effects correlations. The disadvantages of the SEBLUP are that it is not robust to outliers and post-design consistency and does not take into account the possibility of spatial non-stationarity (Pratesi, 2015, pp. 42–43). A generalisation of this predictor that takes into account local non-stationarity of spatial dependence is considered in the work of Benedetti et al. (2012).

The mean squared error of the predictor (3.50) under the assumption of normality of random effects is given by the following formula (Pratesi and Salvati, 2008, p. 117):

$$MSE\left(\tilde{\theta}_d\left(\hat{\sigma}_u^2, \hat{\rho}\right)\right) = g_{1d}\left(\sigma_u^2, \rho\right) + g_{2d}\left(\sigma_u^2, \rho\right) + E\left(\tilde{\theta}_d\left(\hat{\sigma}_u^2, \hat{\rho}\right) - \tilde{\theta}_d\left(\sigma_u^2, \rho\right)\right)^2, \tag{3.51}$$

where (Pratesi and Salvati, 2008, p. 136):

$$g_{1d}\left(\sigma_u^2,\rho\right) = \mathbf{b}_d^T\left(\mathbf{G}-\mathbf{G}\mathbf{Z}^T\left(\mathbf{R}+\mathbf{Z}\mathbf{G}\mathbf{Z}^T\right)\mathbf{Z}\mathbf{G}\right)\mathbf{b}_d, \tag{3.52}$$

$$\begin{aligned} g_{2d}\left(\sigma_u^2,\rho\right) = {} & \left(\mathbf{x}_d-\mathbf{b}_d^T\mathbf{G}\mathbf{Z}^T\left(\mathbf{R}+\mathbf{Z}\mathbf{G}\mathbf{Z}^T\right)^{-1}\mathbf{X}\right)\times \\ & \times\left(\mathbf{X}^T\left(\mathbf{R}+\mathbf{Z}\mathbf{G}\mathbf{Z}^T\right)^{-1}\mathbf{X}\right)^{-1}\times\left(\mathbf{x}_d-\mathbf{b}_d^T\mathbf{G}\mathbf{Z}^T\left(\mathbf{R}+\mathbf{Z}\mathbf{G}\mathbf{Z}^T\right)^{-1}\mathbf{X}\right)^T. \end{aligned} \tag{3.53}$$

The last component (3.51) can be approximated using Taylor expansion. It should be added that this component, also denoted as $g_{3d}\left(\sigma_u^2,\rho\right)$, results from the estimation of the variance components. Following Rao (2003), however, the component $g_1\left(\sigma_u^2,\rho\right)$ results from the random effects estimation, and the $g_2\left(\sigma_u^2,\rho\right)$ from estimation of $\boldsymbol{\beta}$. These components are of order $O(1)$ and $O(D^{-1})$, respectively.

Estimators based on parametric as well as non-parametric bootstrap methods, among others, can be used to estimate the mean squared error for the SEBLUP. In their paper, Molina et al. (2009) present the possibility of using the estimator considered by González-Manteiga et al. (2008) for MSE estimation, which is discussed in more detail in subsection 3.4.

The applicability of the Spatial EBLUP under the assumption of the Fay–Herriot model in small area estimation was presented by Petrucci et al. (2005). The authors used this predictor to estimate average olive production on farms in Tuscany. The problem also addresses the economic aspects of agriculture. The analyses used data from The Farm Structure Survey.

For the SEBLUP for type B models, which include the model of Battese et al. (1988), given by the formula (1.94), the vector of correlated random effects $\mathbf{v}$ has the same form as for type A models, and hence is given by the formula (3.46). It should be added, following Pratesi (2015), that the matrix $\mathbf{I}-\rho\mathbf{W}$ should be positively defined in order to determine the inverse matrix. This condition is fulfilled when $\rho \in \left(\frac{1}{\max_i(\lambda_i)},\frac{1}{\min_i(\lambda_i)}\right)$, where λ_i are the eigenvalues of the matrix $\mathbf{W}$.

Similar to the Spatial EBLUP for type B models, it requires individual data for both the study variable and auxiliary variables for the sampled elements and the values of the auxiliary variables for the out-of-sample elements. A matrix of spatial weights for the drawn and out-of--sample domains is also required. Predictors of this class are also – like the EBLUP – sensitive to outliers. This problem has been addressed by Schmid and Münnich (2014), among others.

3.5.2. REBLUP

The problem of sensitivity to outliers of model parameter estimators also used in prediction using the EBLUP is an issue that has been frequently addressed in the literature. This topic in the context of parameter estimations obtained using the generalised least squares method, the maximum likelihood method or the restricted maximum likelihood method was considered, among others, in papers by Stahel and Welsh (1992), Huggins (1993), and Richardson (1997). It should be added that in the papers, the considerations were based on linear mixed models.

Sinha and Rao (2009) proposed a generalisation of the EBLUP based on a general linear mixed model with a block-diagonal covariance matrix to address the robustness of the EBLUP to outliers (Robust EBLUP). In what follows, the authors also consider a special case of this model, a model with a nested random component given by the formula (1.94) and belonging to type B of models.

As emphasised in their paper by Schmid and Münnich (2014), in this case the modification of the density function of the considered random variable due to $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ is maximised. This makes it possible to obtain the robust maximum-likelihood equation proposed by Richardson and Welsh (1995), the solution of which is the robust estimators of the above quantities. It should be added that the Newton-Raphson algorithm can be used to solve the maximum likelihood equation. Parameter estimates are thus obtained based on the following equations (Sinha and Rao, 2009, p. 384):

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{U}^{1/2}\boldsymbol{\psi}(\mathbf{r}) = 0, \tag{3.54}$$

$$\boldsymbol{\psi}(\mathbf{r})\mathbf{U}^{1/2}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \Theta_l}\mathbf{V}^{-1}\mathbf{U}^{1/2}\boldsymbol{\psi}(\mathbf{r}) - \operatorname{tr}\left[\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \delta_l}\mathbf{K}\right] = 0, \tag{3.55}$$

where $\mathbf{r} = \mathbf{U}^{1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is the unit-level vector of residuals, and $\mathbf{U}^{1/2}$ is the root of the diagonal matrix containing the diagonal elements of the variance-covariance matrix $\mathbf{V}$, the matrix $\mathbf{K}$ can be written as:

$$\mathbf{K} = E\left[\boldsymbol{\psi}^2(\mathbf{r})\right], \tag{3.56}$$

where $\boldsymbol{\psi}$ is the Huber (1964) function. The l-th variance component is denoted by δ_l. In addition, the authors proposed to use Fellner's (1986) equation for the estimation of the random effects $\mathbf{v}$:

$$\mathbf{Z}^T\mathbf{R}^{-1/2}\psi\left(\mathbf{R}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v})\right)\mathbf{G}^{-1/2}\psi\left(\mathbf{R}^{-1/2}\mathbf{v}\right) = \mathbf{0}. \tag{3.57}$$

It should be noted that it is necessary to determine the element of the $\mathbf{R}^{-1}$ matrix. As Sinha and Sattar (2015) point out in their paper, the estimate of $\boldsymbol{\beta}$ is a consistent estimator and has an asymptotic normal distribution. Furthermore, the authors add that the estimate of the variance components of $\boldsymbol{\delta}$ can be obtained using the REML method.

The robust domain mean value predictor for the model (1.89) obtained using (3.54), (3.55) and (3.57) then has the following form (Sinha and Rao, 2009, p. 387):

$$\hat{\Theta}_i^{REBLUP} = N_d^{-1}\left(\sum_{i\in s_d} y_{di} + \sum_{i\in\Omega_d}\left(\mathbf{x}_{di}^T\hat{\boldsymbol{\beta}}_M + \hat{\mathbf{v}}_{dM}\right)\right), \tag{3.58}$$

where $\hat{\boldsymbol{\beta}}_M$ and $\hat{\mathbf{v}}_{dM}$ denote robust fixed-effects and random-effects estimates. It should be noted, however, that the predictor proposed by Sinha and Rao (2009) assumes a block-diagonal variance-covariance matrix. Furthermore, as reported by Chambers et al. (2014), this predictor is based on the assumption that the mean of the random components in the non-sampled units in the study domain converges to 0. As given in Rao et al. (2014), however, it is also possible to use the REBLUP under weaker fixed effects assumptions than linear regression. The authors replaced this assumption with semiparametric regression.

Sinha and Rao (2009) also proposed an estimator of the mean squared error of the predictor (3.58). This estimator is based on the parametric double bootstrap method considered in Hall and Maiti (2006) and described in the fourth subsection of this paper. This method is discussed in more detail in the third subsection of this chapter.

In their paper, Sinha and Rao (2009) also considered the use of the proposed REBLUP class to predict the average area planted with corn and soybeans in twelve counties in Iowa. They used a dataset from the work of Battese et al. (1988). Sinha and Rao (2009) compared the properties of the EBLUP and a proposed robust modification of it.

3.5.3. SREBLUP

The Spatial Robust EBLUP (SREBLUP) proposed by Schmid and Münnich (2014) combines the concepts of a robust predictor and one that takes into account the correlation of random effects using an autoregressive SAR model. This predictor is a modification of the REBLUP of Sinha and Rao (2009), and is therefore based on a type B model, which includes the model of Battese et al. (1988) given by formula (1.94). It should be added that in their work, they only considered a SAR-type model for reasons of practical applications. The random effects vector $\mathbf{v}$, as in the SEBLUP case, is given by the formula (3.45) and the matrix $\mathbf{V}$ by the formula (3.48), where the matrices $\mathbf{G}$ and $\mathbf{R}$ are given by the formulas (3.47) and (3.46), respectively. However, Schmid and Münnich (2014) following Sinha and Rao (2009) suggest determining the estimates of $\boldsymbol{\beta}$, ρ and the variance of both effects and random components using the robust maximum likelihood method (Richardson and Welsh, 1995) and the Newton-Raphson algorithm. This allows the estimation of random effects using the Fellner (1986) equation given by the formula (3.57).

The robust EBLUP, taking into account the spatial correlation of random effects for the domain mean value, is then of the form:

$$\hat{\theta}_d^{SREBLUP} = N_d^{-1}\left(\sum_{i\in s_d} y_{di} + \sum_{j\in\Omega_{rd}}\left(\mathbf{x}_{dj}^T\hat{\boldsymbol{\beta}}^{SR} + z_{dj}\hat{\mathbf{v}}_d^{SR}\right)\right), \tag{3.59}$$

where $d = 1,\ldots,D$, y_{di} are the values of the study variable for the d-th domain elements that were sampled, and $\mathbf{x}_{dj}^T$ and z_{dj} are the values of the auxiliary variables for the elements that were not drawn into the sample. The $\hat{\boldsymbol{\beta}}^{SR}$ and $\hat{\mathbf{v}}_d^{SR}$ are denoted as fixed-effects and random-effects estimates using the robust maximum-likelihood method.

Further modifications of this predictor can also be found in the literature, among others, taking into account representative outliers. These are understood as outliers in the sample for which it is assumed that they have been correctly observed and that the population contains other values similar to them (cf. Schmid et al., 2016). As with the REBLUP for its spatial generalisation of (3.59), it is not possible, due to its high complexity, to write an explicit formula for the mean squared error. Schmid and Münnich (2014) in their paper, however, presented the possibility of estimating it using a parametric bootstrap method. This approach was also considered by Sinha and Rao (2009) for the REBLUP.

As Schmid and Münnich (2014) point out, the SREBLUP can find application for economic data. In a simulation study based on artificial data, the authors compared the properties of the proposed predictor with those of the EBLUP, REBLUP and GREG-type estimators, considering, among other things, relative bias and relative RMSE.

3.5.4. GWEBLUP and RGWEBLUP

Both the EBLUP and SEBLUP assume that the regression coefficients are spatially invariant (spatially stationary). The problem of spatial stationarity was addressed in the work of Beenstock and Felsenstein (2008), among others. This means that the relationship between the study variable and the auxiliary variables is the same for the entire target area. However, the assumption of the same linear correlation across the entire population may not be met in practice, as was considered, among others, in the work of Opsomer et al. (2008), Chandra et al. (2012b), Salvati et al. (2012b), and Chambers et al. (2016). Following Baldermann et al. (2018), one solution to this problem may be to assume local linear, space non-stationary models for the study variable. Included in the prediction process is the Geographically Weighted Regression (GWR) considered in Brunsdon et al. (1996). The GWR is one of the popular spatial interpolation methods designed to spatially interpolate a single data set. These methods allow values of variables at unobserved locations in geographic space to be predicted based on values from observed locations. In the work of Chandra et al. (2012b) the authors extended this by combining the GWR and linear mixed models in predicting characteristics for small areas.

Let u be the geographical coordinates of the location of any individual in the population. A location-sensitive linear mixed model can be written as (cf. Baldermann et al., 2018, p. 141;

Chandra et al., 2012b, p. 2877):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_u + \mathbf{Z}\mathbf{v} + \mathbf{W}_u^{-\frac{1}{2}}\mathbf{e}, \tag{3.60}$$

where **X** and **Z** are the auxiliary variable matrices, $\boldsymbol{\beta}_u$ is the p-element vector of location-specific regression parameters u, **v** and **e** are the vectors of random effects and random components, respectively, and $\mathbf{W}_u$ is the diagonal matrix of geographic weights. The diagonal elements of the $\mathbf{W}_u$ matrix, denoted by $w_i(u)$, are a function of the distance of the i-th population element from location u. Their values decrease as the distance between the i-th population element and location u increases. Chandra et al. (2012b) used the Euclidean weighting function in their analysis to define geographic weights. A comprehensive review of the different weighting functions in the GWR can be found in Fotheringham et al. (2002). It should be noted that this approach assumes that there are as many superpopulation models as there are individuals in the population under study, due to the dependence of the model (3.60) on the location u.

A modification of the EBLUP assuming the model (3.60) is referred to in the literature as the geographically weighted EBLUP – GWEBLUP. When the estimated parameter is the mean value, it has the following form (Baldermann et al., 2018, p. 142):

$$\begin{aligned}\hat{\theta}_d^{GWEBLUP} &= N_d^{-1}\left(\sum_{i\in s_d} y_{di} + \sum_{i\in\Omega_{rd}} \hat{y}_{dj}^{gw}\right)\\ &= N_d^{-1}\left(\sum_{i\in s_d} y_{di} + \sum_{j\in r_d}\left(x_{dj}^T(\hat{\beta}_{dj}^{gw})^T + (\hat{v}_d^{gw})^T\right)\right),\end{aligned} \tag{3.61}$$

where $d = 1,\dots,D$, $i = 1,\dots,n_d$ and $j = 1,\dots,N_r$, $\hat{\beta}_{dj}$, and $\hat{v}_d^T$ are estimators obtained using geographically weighted regression. The iterative algorithm that can be used to determine them is discussed in more detail in Chandra et al. (2012). It should be noted that there are parameters estimated for all in-sample and out-of-sample units. It implies large computational requirements.

The conditional MSE estimator of the predictor (3.61) uses the pseudolinearisation approach proposed by Chambers et al. (2011). Chandra et al. (2012b) also proposed a modification of this estimator based on an extension of the approach proposed in the work of Chambers et al. (2011).

When making predictions under the assumption of the model (3.60), it should be noted that observations that are far away from location u are assigned smaller weights compared to observations that are closer. The implication is that extreme observations have little impact on the parameter estimates, compared to values that are at a short distance from location u. As noted by Baldermann et al. (2018) in the first case, therefore, limiting the effect of outliers on the estimates may not be necessary, as already taking geographical weights into account may allow robust estimates to be obtained. In the second case, and therefore outliers close to the location of u, this limitation becomes crucial, as the inclusion of geographical weights may amplify the effect of outliers. The authors proposed a modification to the GWEBLUP that takes into account both geographic weights and an influence function to reduce the effect of outliers on the parameter estimate.

The prediction procedure using the RGWEBLUP, and thus the Robust GWEBLUP, is based on the work of Sinha and Rao (2009), who proposed a robust modification of the EBLUP. The first step to derive a robust predictor is to maximise the reliability function of the variable under test against $\boldsymbol{\delta}$ and the local coefficients of $\boldsymbol{\beta}_{dj}$ by solving the robust ML equations (Sinha and Rao, 2009). In the next step, the robust estimators are used to estimate the random effects of $\mathbf{v}$ according to the method proposed by Fellner (1986).

The robust GWEBLUP is obtained by substituting in (3.61) the estimators $\hat{\boldsymbol{\beta}}_{di}^{gw}$ and $\hat{\mathbf{v}}_{d}^{gw}$ with their robust counterparts $\hat{\boldsymbol{\beta}}_{di}^{\psi,gw}$ and $\hat{\mathbf{v}}_{d}^{\psi,gw}$. Following Baldermann et al. (2018), it should be added that it is not possible to represent both $\hat{\boldsymbol{\beta}}_{di}^{\psi,gw}$ and $\hat{\mathbf{v}}_{d}^{\psi,gw}$ using an explicit formula.

The conditional MSE estimator for the RGWEBLUP can be obtained based on the full linearisation approach proposed by Chambers et al. (2014). In determining the conditional mean squared error, we treat the random effects as fixed but unknown. The idea of the linearisation approach is to decompose the MSE into the variance of the forecast error and the bias quadratic. Given this decomposition, the MSE estimator for the RGWEBLUP of the domain mean value can be written as:

$$\widehat{MSE}\left(\hat{\theta}_{d}^{RGWEBLUP}\right) = h_{1d}\left(\hat{\boldsymbol{\beta}}_{dj}^{\psi,gw}, \hat{\mathbf{v}}_{d}^{\psi,gw}\right) + h_{2d}\left(\hat{\boldsymbol{\beta}}_{dj}^{\psi,gw}, \hat{\mathbf{v}}_{d}^{\psi,gw}\right) + \\ + h_{3d}\left(\hat{\boldsymbol{\beta}}_{dj}^{\psi,gw}, \hat{\mathbf{v}}_{d}^{\psi,gw}\right) + \left[B\hat{ias}\left(\hat{\theta}_{d}^{RGWEBLUP}\right)\right]^{2}, \quad (3.62)$$

where h_{1d} is the component accounting for variation due to estimation of regression coefficients and area-specific random effects, h_{2d} is the residual variance, and h_{3d} is the component accounting for variation due to estimation of variance components. The bias of the robust GWEBLUP is $B\hat{ias}\left(\hat{\theta}_{d}^{RGWEBLUP}\right)$.

Chandra et al. (2012b) made a comparison in a simulation study of the properties of the EBLUP and its modification – GWEBLUP. The data used in the analyses by the authors came from the Australian Agricultural and Grazing Industries Survey (AAGIS) conducted by the Australian Bureau of Agricultural and Resource Economics. The study addressed the problem of estimating MSEs for prudent predictors, considering the estimators proposed by Chambers et al. (2011) and based on the work of Prasad and Rao (1990).

In their paper, Baldermann et al. (2018) presented the applicability of the RGWEBLUP and its modification incorporating some bias correction to estimate net rent per square metre for residential areas in Berlin. In their analyses, the authors used data from the German real estate market database provided by Empirica-Systeme GmbH from 2015. The authors compared the properties of the considered predictors with the Horvitz-Thompson estimator.

3.5.5. Pseudo-EBLUP

When discussing the problem of prediction using the EBLUP, one should also mention a class of predictors called pseudo empirical best linear unbiased predictors. Among the first papers to consider predictors belonging to this class are the articles by Prasad and Rao (1999) and You and Rao (2002). Pseudo-EBLUPs, following Graf et al. (2019), are one method that allows the introduction of weights derived from the sampling method into the estimation procedure based on a model-based approach.

Prasad and Rao (1999) in their paper proposed a pseudo-EBLUP for the mean value in the domain θ_d, using the p-consistent estimator $\bar{y}_{dw}$. The proposed pseudo-EBLUP assumes an aggregate type A model having the form (cf. You and Rao, 2002, p. 433):

$$\bar{y}_{dw} = \bar{x}_{dw}^T \beta + v_d + \bar{e}_{dw} \tag{3.63}$$

for $d = 1, \ldots, D$, where $\bar{x}_{dw} = \sum_{i=1}^{n_d} w_{di} x_{di}$ and $\bar{e}_{dw} = \sum_{i=1}^{n_d} w_{di} e_{di}$ with an expected value of 0 and a variance $\sigma_e^2 \sum_{i=1}^{n_d} w_{di}^2 \equiv \sigma_e^2 \delta_d^2$. It should be added that the model (3.63) is formed by combining the estimator given by the formula:

$$\bar{y}_{dw} = \frac{\sum_{i=1}^{n_d} \tilde{w}_{di} y_{di}}{\sum_{i=1}^{n_d} \tilde{w}_{di}} = \sum_{i=1}^{n_d} w_{di} y_{di} \tag{3.64}$$

and a model with a nested random component (1.94). In addition, for unit-level weights $\tilde{w}_{di}$, there is $w_{di} = \frac{\tilde{w}_{di}}{\sum_{i=1}^{n_d} \tilde{w}_{di}}$ and $\sum_{i=1}^{n_d} w_{di} = 1$. The BLUP of the domain mean value is thus given by the formula (cf. You and Rao, 2002, p. 434):

$$\tilde{\theta}_{d,aw} = \gamma_{dw} \bar{y}_{dw} + \left(\bar{X}_d - \gamma_{dw} \bar{x}_{dw}\right)^T \tilde{\beta}_{aw}, \tag{3.65}$$

where $\gamma_{dw} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2 \delta_d^2}$, $\tilde{\beta}_{aw} = \left(\sum_{d=1}^{D} \gamma_{dw} \bar{x}_{dw} \bar{x}_{dw}^T\right)^{-1} \left(\sum_{d=1}^{D} \gamma_{dw} \bar{x}_{dw} \bar{y}_{dw}\right)$, $\bar{X}_d$ – vector of mean values of auxiliary variables in the population, and $\bar{y}_{dw}$ and $\bar{x}_{dw}$ – mean values of the sampled and auxiliary variables. You and Rao (2002) point out that the use of the $\tilde{\beta}_{aw}$ estimator based on the aggregate model (3.63) may result in a loss of efficiency compared to estimates based on the unit-level, type B model, and therefore also a loss of efficiency in estimating the analysed characteristic in the domain.

Substituting in the (3.65) σ_e^2 and σ_d^2 their estimates obtained by the method considered in Nair's (1941) paper – methods of fitting-of-constants, which are given by formulae (You and

Rao, 2002, p. 433):

$$\hat{\sigma}_e^2 = \frac{1}{n-D-p+1}\sum_{d=1}^{D}\sum_{i=1}^{n_d}\hat{\varepsilon}_{di}^2 \tag{3.66}$$

and

$$\hat{\sigma}_v^2 = \max\left(\tilde{\sigma}_v^2, 0\right), \qquad \tilde{\sigma}_v^2 = \frac{1}{n^*}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}\hat{u}_{di}^2 - (n-p)\hat{\sigma}_e^2\right) \tag{3.67}$$

we obtain a pseudo-EBLUP of the mean value in the domain. It is important to add that $\hat{\varepsilon}_{di}^2$ and $\hat{u}_{di}^2$ are the residuals of the model estimated by the classical least squares method for the considered variables and for the adjusted variables, for which the values are determined as the differences of the original values and the mean value in the domain (e.g. $y_{id} - \hat{y}_d$). In addition, $n^* = n - \text{tr}\left(\left(X^T X\right)^{-1}\sum_{d=1}^{D} n_d^2 \bar{x}_d \bar{x}_d^T\right)$, and p denotes the number of auxiliary variables.

As described in Prasad and Rao (1990), the pseudo-EBLUP mean squared error can be approximated by the following formula:

$$MSE\left(\tilde{\theta}_{d,aw}\right) \approx g_{1dw}\left(\sigma_e^2, \sigma_v^2\right) + g_{2dw}\left(\sigma_e^2, \sigma_v^2\right) + g_{3dw}\left(\sigma_e^2, \sigma_v^2\right), \tag{3.68}$$

where

$$\begin{aligned}
g_{1dw}\left(\sigma_e^2, \sigma_v^2\right) &= \left(1-\gamma_{dw}\right)\sigma_v^2, \\
g_{2dw}\left(\sigma_e^2, \sigma_v^2\right) &= \left(\bar{X}_d - \gamma_{dw}\hat{x}_{dw}\right)^T \Phi_{aw}\left(\bar{X}_d - \gamma_{dw}\hat{x}_{dw}\right), \\
g_{3dw}\left(\sigma_e^2, \sigma_v^2\right) &= \gamma_{dw}\left(1-\gamma_{dw}\right)^2 \sigma_e^{-4}\sigma_v^{-2} h\left(\sigma_e^2, \sigma_v^2\right), \\
h\left(\sigma_e^2, \sigma_v^2\right) &= \sigma_e^4 var\left(\tilde{\sigma}_v^2\right) - 2\sigma_e^2\sigma_v^2 cov\left(\hat{\sigma}_e^2, \tilde{\sigma}_v^2\right) + \sigma_v^4 var\left(\hat{\sigma}_e^2\right),
\end{aligned}$$

and Φ_{aw} denotes the covariance matrix $\mathbf{Y}$. The following estimator can be used to estimate the mean squared error of the predictor proposed by Prasad and Rao (1999):

$$\widehat{MSE}\left(\tilde{\theta}_{d,aw}\right) = g_{1dw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{2dw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{3dw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right), \tag{3.69}$$

where $g_{1dw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$, $g_{2dw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$, and $g_{3dw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$ are determined as in the case of (3.68) for $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$.

It should be added that the authors in their analyses used data considered also by Battese et al. (1988). Jiang and Lahiri (2006) mention the problem of generalising this predictor to models, such as the generalised linear mixed model (GLMM), as a drawback. The difficulty of its generalisation, the authors state, is due to the very complex assumptions of this type of model for estimators consistent after the sampling design. Furthermore, the model considered by Prasad and Rao (1990), as reported by Jiang and Lahiri (2006), is clearly not suitable for complex sampling designs, such as multi-stage stratified sampling.

In the case of the pseudo-BLUP proposed by You and Rao (2002), we also assume an aggregate model (3.63). Furthermore, the authors assumed that the parameters of this model, i.e. $\boldsymbol{\beta}$, σ_e^2, and σ_v^2, are known. In this case, the pseudo-BLUP of the domain mean value is given by the formula (You and Rao, 2002, p. 435):

$$\tilde{\theta}_{dw} = \gamma_{dw}\bar{y}_{dw} + \left(\bar{X}_d - \gamma_{dw}\bar{x}_{dw}\right)^T \boldsymbol{\beta}. \tag{3.70}$$

The σ_e^2 and σ_v^2 estimates are obtained based on the type B model and, therefore, according to the formulae (3.66) and (3.67). A previous estimator of the random effects estimator based on the aggregate model (3.63) is required to determine the estimation of $\boldsymbol{\beta}$ (You and Rao, 2002, p. 435):

$$\tilde{v}_{dw}\left(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2\right) = \gamma_{dw}\left(\bar{y}_{dw} - \bar{x}_{dw}^T\boldsymbol{\beta}\right) \tag{3.71}$$

and solving the equation:

$$\sum_{d=1}^{m}\sum_{i=1}^{n_d} \tilde{w}_{di}x_{di}\left(y_{di} - x_{di}^T\boldsymbol{\beta} - \tilde{v}_{dw}\left(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2\right)\right) = 0. \tag{3.72}$$

The resulting $\boldsymbol{\beta}$ estimator has the following form:

$$\tilde{\boldsymbol{\beta}}_w = \left(\sum_{d=1}^{m}\sum_{i=1}^{n_d} \tilde{w}_{di}x_{di}\left(x_{di} - \gamma_{dw}\bar{x}_{dw}\right)^T\right)^{-1}\left(\sum_{d=1}^{m}\sum_{i=1}^{n_d} \tilde{w}_{di}\left(x_{di} - \gamma_{dw}\bar{x}_{dw}\right)y_{di}\right). \tag{3.73}$$

Thus, by substituting (3.66), (3.66), and (3.73) into (3.70) we obtain the pseudo-EBLUP given by the formula (You and Rao, 2002, p. 436):

$$\hat{\theta}_{dw} = \hat{\gamma}_{dw}\bar{y}_{dw} + \left(\bar{X}_d - \hat{\gamma}_{dw}\bar{x}_{dw}\right)^T \hat{\boldsymbol{\beta}}_w. \tag{3.74}$$

Similar to the pseudo-EBLUP proposed by Prasad and Rao (1990), it is possible to approximate the mean squared error according to the formula (3.68). The estimator presented in the paper by Prasad and Rao (1990), given by the formula (3.69), can also be used to estimate the MSE. It should be added that the bias of this estimator is of the order of $o\left(D^{-1}\right)$. In Torabi and Rao (2010), the authors also considered the possibility of using the double bootstrap method in MSE estimation proposed by Hall and Maiti (2006).

The predictor proposed by You and Rao (2002) also allows for inter-domain borrowing of power from the model as well as the inclusion of weights to preserve the p-consistency of the predictor. It should be noted, however, that unlike the predictor considered by Prasad and Rao (1999), the parameter estimation process takes into account not only the aggregate model (the type A model) which is a combination of the model belonging to type B and the direct estimator. The authors also include the type B model itself and the weights in the parameter estimation process. The same set of variables is therefore required for the model and the direct estimator. The predictor considered by You and Rao (2002) also has, unlike the pseudo-EBLUP proposed by Prasad and Rao (1999), the property of self-benchmarking. That is, the sum of the estimates of the total value for small areas equals the direct estimate of the regression estimator of the total value for the large domain.

In their paper, You and Rao (2002) also presented an example of the application of the pseudo-EBLUP in small area estimation in an issue in agricultural economics. In their analyses, the authors considered the prediction of average corn and soybean acreage per segment for counties in north-central Iowa, based on data from the work of Battese et al. (1988). The authors compared the properties of the proposed pseudo-EBLUP with the EBLUP and the pseudo--EBLUP presented in the paper by Prasad and Rao (1999).

3.5.6. NPEBLUP

The generalisation of the EBLUP proposed in the work of Opsomer et al. (2008) (Non--Parametric Empirical Best Linear Unbiased Predictor – NPEBLUP) allows prediction when the dependence between the test variable and the auxiliary variables is more complex than in the case of a linear model. It should be added that the authors included both random effects and a smoothed non-parametric trend in the considered type B model. In the simplest case, the model analysed has the following form (cf. Opsomer et al., 2008, p. 267):

$$Y_{di} = m\left(x_{di}\right) + z_{di} v_d + e_{di} \quad \left(i = 1, \ldots, n_d;\ d = 1, \ldots, D\right), \tag{3.75}$$

where $m(.)$ is the unknown smoothing function of the auxiliary variable x_{di}, and z_{di} is a constant whose values are known for all individuals in the population. It should be added that we assume that the function $m(.)$ can be approximated sufficiently well by a P-spline function. In addition, we assume that the effects and random components have a normal distribution with an expectation value of 0 and a variance of σ_v^2 and σ_e^2, respectively.

The nonparametric EBLUP of the mean value in the domain for the model (3.75) is given by the formula (Pratesi, 2015, p. 64):

$$\hat{\theta}_i^{NPEBLUP} = N_d^{-1} \left(\sum_{i \in s_d} y_{di} + \sum_{i \in \Omega_{rd}} \hat{y}_{di} \right), \tag{3.76}$$

where $\hat{y}_{di} = \hat{m}\left(x_{di}\right) + z_{di} \hat{\mathbf{v}}_{di}$. It should be added that $\hat{m}$ and $\hat{\mathbf{v}}_{di}$ are estimators obtained using the smoothing function of the auxiliary variable and the estimate of random effects.

It should be noted that the NPEBLUP can allow spatial dependencies to be captured using penalised spline functions. This can be important when the functional form of the dependence between variables is not specified and the dataset under consideration is characterised by complex patterns of spatial dependence (Pratesi, 2015, p. 86). When assessing the NPEBLUP mean squared error, it is possible to use estimators based on both parametric and non-parametric bootstrap methods, which are discussed in more detail in the previous section of this book.

In the paper by Opsomer et al. (2008), the use of the proposed predictor in environmental studies was also presented. The analyses considered data collected from the US Enviromental Protection Agency's Environmental Monitoring and Assessment Programme. The paper presents the problem of predicting the average Acid Neutralising Capacity of lakes in the northeastern US. The prediction problem using the NPEBLUP was also considered in the paper by Ugarte et al. (2009).

3.6. EBLUP applications

This subsection will present selected applications of EBLUP in small area estimation of an economic nature. The EBLUPs are used in many areas, including quality of life and poverty analyses, corporate finance analyses, agricultural economics, economic aspects of health policy, ecology and transport policy.

Issues in the area of quality of life and poverty have been considered, among others, by Pratesi and Salvati (2008). This paper addresses the problem of predicting average income per capita at the level of Tuscan sub-regions using the EBLUP and some modification of it. The research used, among others, data from the Italian Decennial Census of Population, databases of the Istituto Regionale Programmazione Economica and administrative records. In the paper by Jędrzejczak and Kubacki (2016), the authors applied the EBLUP to predict the average value of disposable income and self-employment income in Polish voivodeships based on the Fay–Herriot model and the Rao–You model, 1992. The analyses used data from the Household Budget Survey and the Local Data Bank. Namazi-Rad and Steel (2015), in their paper, addressed the problem of model selection in the context of type A and B models. The analyses were conducted using artificial data generated from Australian Census information. Chandra et al. (2018) considered the EBLUP for the Fay–Herriot model and in their research compared its properties with empirical plug-in predictors (EPPs), among others. The aim of the study was to apply the aforementioned methods to predict the percentage of indebted households using data from investment and debt surveys conducted in India. This issue was also addressed in the work of Chandra et al. (2010). The authors used the EBLUP for data from the Debt-Investment Survey conducted by the National Sample Survey Organisation. The authors considered the prediction of the average amount of loans taken out by a household and outstanding.

Applications of the EBLUP in the field of agricultural economics can be found, among others, in the paper by Esteban et al. (2012), which addresses the prediction of the percentage of households in poverty. The data used in the analyses came from the Spanish Living Conditions

Survey (SLCS). Rivest et al. (2016) also considered the problem of predicting crop area, but based on data analysed by Battese et al. (1988). The authors made a comparison between the properties of the EBLUP and Empirical Best Unbiased Predictors (EBUP). Millitino et al. (2006) used the EBLUP to predict the total area occupied by olive trees in the Comarca IV area in Spain. An article by Jiang and Nguyen (2012) addressed the problem of predicting the average area planted with corn in counties in north-central Iowa using the EBLUP. The data considered were presented in the paper by Battese et al. (1988). In their discussion, the authors compared the properties of the EBLUP and some modification of it.

The application of the EBLUP in the context of corporate finance analysis is presented, for example, in the work of Ghosh and Rao (1994). The authors addressed the problem of predicting corporate remuneration using the EBLUP. The authors conducted their considerations based on an artificial population intended to resemble the dataset analyzed by Särndal and Hidroglou (1989). In addition, Ghosh and Rao (1994) compared the properties of the EBLUP with selected estimators and predictors.

The problem of applying the EBLUP in the area of health policy more precisely financing of medical facilities, can be found in the work of Jiang and Tang (2011), among others. The authors considered the EBLUP for the Fay–Herriot model. The study used data considered in the work of Morris and Christiansen (1996). The issues were also addressed in the paper by Jiang et al. (2011). However, the authors used artificial data. The paper compares the properties of the EBLUP considering four different estimators of random effects variance.

Mauro et al. (2016) considered the problem of prediction using the EBLUP, in the context of forest resources and environmental protection. The data concerned a pine forest located in the Valle de las Caderechas, Spain. Chambers et al. (2014) also addressed an issue from the field of ecology. The authors considered the problem of using the EBLUP to predict the acid neutralising capacity of lakes. The data included in the analysis were obtained from the US Environmental Protection Agency's Environmental Monitoring and Assessment Program (EMAP). It should be added that the authors compared the properties of the EBLUP with some modification in their simulation studies. The paper by Petrucci and Salvati (2006), similarly to the paper by Pratesi and Salvati (2008), considered a modification of the EBLUP taking into account spatial correlation, however, this predictor was used to assess average erosion in the Rathbun Lake catchment area in the southern part of Iowa.

Hall and Maiti (2006) considered an issue of potential relevance to transport policy in their paper. The authors approached the problem of using the bootstrap method in the estimation of

EBLUP mean squared error and prediction intervals. The prediction procedure used data analysed by Nusser and Goebel (1997) or Wang and Fuller (2003), among others.

3.7. Summary

In this chapter, the issues of prediction using BLUPs and EBLUPs were addressed. In the first section, the predictors proposed by Henderson (1950) and Royall (1976) were discussed. For each of these predictors, the assumptions that are made when predicting using them were discussed, with particular reference to the prediction of characteristics in the domain. The form of the mean squared error of the predictors considered are also presented in this subsection. Selected applications of the predictors proposed by Henderson (1950) and Royall (1976) were also discussed.

The second subsection was focused on the EBLUPs in the light of the classification of linear mixed models into type A and B models. Again, the assumptions made for each of the predictors considered and the form of the mean squared error of the prediction were presented. Both sections of this subsection conclude with examples of the application of the predictors discussed. The third subsection proposed the use of EBLUPs assuming the special case of a linear mixed model with correlated random effects vectors.

The next section dealt with the issue of estimating the mean squared error of EBLUPs. Among others, the classical estimator was presented, as well as estimators based on the parametric bootstrap method or the jackknife method. The subsection also presents the properties of the presented estimators, and their advantages and disadvantages.

In the fifth subsection, selected modifications of the EBLUP were presented. Among the discussed EBLU modifications are the robust predictor (REBLUP), the predictor taking into account the spatial correlation of random effects (SEBLUP), and the combination of the two mentioned modifications (SREBLUP). Among the presented EBLUP modifications, there was also a geographically weighted predictor (GWEBLUP) with its robust variant (RGWEBLUP), a pseudo-EBLUP, and a non-parametric EBLUP. It should be added that the advantages and disadvantages of the presented predictors and examples of their application are also discussed in this chapter.

The final subsection provides an overview of selected applications of the economic predictors considered in this chapter. The presented possibilities for the use of EBLUPs include areas such as quality of life and poverty analyses, corporate finance analyses, agricultural economics, economic aspects of health policy, and transport and ecology.

The author’s proposals presented in this chapter include applications of EBLUs assuming linear mixed models with correlated random effects vectors in small area estimation with an example. They also deal with modifications of known methods for estimating the mean squared prediction error that allow estimating the accuracy of the proposed EBLUs.

Chapter 4

Empirical best predictors and plug-in predictors

This chapter will discuss two further classes of predictors – empirical best predictors and plug-in predictors. The problem of prediction using these two classes of predictors, assuming the special cases of mixed models with correlated random effects presented in subsection 1.2.3, will also be proposed. These issues will also be complemented by estimates of the mean squared error of the empirical best predictors and plug-in predictors, as well as their application in studies of an economic nature.

4.1. Empirical best predictor

When considering the problem of predicting any function of random variables Y, denoted as θ, we assume the decompositions of the vector $\mathbf{Y}$ and the matrix of auxiliary variables $\mathbf{X}$ given by the formulas (3.1) and (3.2). Among the predictors $\hat{\theta}$ of functions of random variables θ, the best predictor (BP) is the one that minimises the mean squared error, thus (cf. Molina and Rao, 2010, p. 372):

$$MSE(\hat{\theta}) = E_{\xi}(\hat{\theta} - \theta)^2. \tag{4.1}$$

Hence, the best predictor is given by the formula:

$$\hat{\theta}_{BP} = E(\theta|\mathbf{Y}_s), \tag{4.2}$$

meaning that it can be determined as the conditional expectation value of the function of the random variables θ, assuming that the form of the conditional distribution $\mathbf{Y_r}|\mathbf{Y_s}$ is known. This distribution in practice depends on a vector of unknown parameters $\boldsymbol{\tau}$. In the case of the general mixed model given by the formula (1.89), it depends on the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, thus the vectors of the fixed effects parameters and the variance components. If the vector $\boldsymbol{\tau}$ is replaced by its estimator, we obtain the Empirical Best Predictor (EBP) denoted as $\hat{\theta}_{EBP}$.

The value of the empirical best predictor of any function of random variables $\theta(\mathbf{Y})$ can be obtained using Monte Carlo approximation. This procedure can be divided into four steps.

Algorithm 2. Monte Carlo EBP approximation

```
1. Estimate the $\boldsymbol{\tau}$ vector of parameters of the distribution of the random variables $\mathbf{Y}$ using the realisation of the $\mathbf{Y_s}$ vector and obtain the $\hat{\boldsymbol{\tau}}$ estimator.
2. Loop for $l = 1$ to $L$; assuming that the form of the distribution $\mathbf{Y_r}|\mathbf{Y_s}$ is known:
   2.1. generate vectors $\mathbf{Y_r}$ ($\mathbf{Y_r^{(l)}}$, $l = 1, 2, \dots, L$), where the vector $\boldsymbol{\tau}$ is replaced by its estimator,
   2.2. construct vectors such that $\mathbf{Y^{(l)}} = \begin{bmatrix} \mathbf{Y_s^T} & \mathbf{Y_r^{(l)T}} \end{bmatrix}^{\mathbf{T}}$, where $l = 1, 2, \dots, L$.
   2.3. End of loop.
3. Calculate the value of the empirical best predictor of the function of random variables $\theta(\mathbf{Y})$:
       $\hat{\theta}_{EBP} = L^{-1} \sum_{l=1}^{L} \theta(\mathbf{Y^{(l)}}).$     (4.3)
```

It should be added, following Molina and Rao (2010), that the realisations of the vector of random variables $\mathbf{Y}$ need not be the values of the study variable, but the values of the study variable after some transformation ($\mathbf{Y} = \mathbf{T}(\mathbf{Y}^*)$), where $\mathbf{Y}^*$ are the values before the analyzed transformation. Then the assumptions about the distribution of ξ are made for the variable after the transformation (e.g. logarithmisation) and it is necessary to take into account the inverse transformation: $\hat{\theta}_{EBP} = L^{-1} \sum_{l=1}^{L} \theta(\mathbf{T}^{-1}(\mathbf{Y^{(l)}}))$.

When addressing the issue of prediction using the EBP, it is important to note the problem of determining the MSE. As reported by Diallo and Rao (2018), the explicit form of the mean squared error of the EBP and its estimators does not exist. The problem of MSE estimation will be discussed further in subsection 4.4.

4.2. Plug-in predictor

The empirical plug-in predictor (EPP) is based on observed values of the study variable for each unit in the sample and estimates for units outside the sample using consistent estimators (Boubeta et al., 2017, p. 37). Ing (2004) notes that this predictor is classified as a multivariate predictor. A plug-in predictor for:

$$\theta = \theta(\mathbf{T}^{-1}(\mathbf{Y})) = \theta\left(\mathbf{T}^{-1}\left(\begin{bmatrix} \mathbf{Y}_s^T & \mathbf{Y}_r^T \end{bmatrix}^T\right)\right)$$

can therefore be written as (cf. Chwila and Żądło, 2019, p. 20):

$$\hat{\theta}_{PLUG-IN} = \theta\left(\mathbf{T}^{-1}\left(\begin{bmatrix} \mathbf{Y}_s^T & \hat{\mathbf{Y}}_r^T \end{bmatrix}^T\right)\right), \tag{4.4}$$

where $\hat{\mathbf{Y}}_r^T$ is a vector of values obtained based on the model assumed for unobserved random variables, where the dependent variable is the post-transformed study variable.

In their paper, Chandra et al. (2018) presented the possibility of using the plug-in method in prediction due to the type of data available and their level of aggregation. The first case is an analysis in which the values of the auxiliary variables are available for the units in the sample, but they are unavailable for the units outside the sample. This is, as the authors point out, the most common situation in many countries where censuses are not regular or censuses are regular but unit-level information is not available. In such cases, Chandra et al. (2018) propose some modification of the EPP for a small area. It uses a synthetic value of the variable of interest that borrows power from the other domains and is therefore based on an increased effective sample size. It is therefore expected to be more effective than a design-based direct estimator (DIR) using only sample data. However, depending on the goodness of fit of the model to the data, this estimator may have a higher bias than DIR. This modification is equivalent to replicating according to the values of the weights (e.g. the inverses of the first-order inclusion probabilities) to replicate the dataset of the auxiliary variables for the whole population. The second case discussed is when the values of the auxiliary variables are available for the individuals in the sample, as well as in aggregated form at the small area level for the population. The aggregated values of auxiliary variables for individual areas are obtained from census or administrative sources. In the last example considered, the values of the auxiliary variables are available as aggregated domain-level values for the population, but are not available for the individuals in the sample. In this case, the predictors of the considered parameter for the small area discussed by the authors cannot be used.

Jiang (2003) points out that the empirical plug-in predictor (EPP) is widely used as an alternative to EBP and GLMM-class models. Chandra et al. (2012b) further note that the EPP predictor can be applied when information on auxiliary variables at the unit or domain level for the population (unit level EPP, area level EPP) are available. It is also possible to apply it to binary data and GLMMs with a logit link function, as considered by Chandra et al. (2012b), Rao (2003), and Saei and Chambers (2003), among others. Morales et al. (2021) point out that empirical best linear unbiased predictors (EBLUP) can be considered plug-in BLUP estimates, which are obtained by replacing the parameters of the unknown variance component vector with consistent estimators. These predictors can therefore asymptotically inherit the desirable good properties of BLUPs.

Pinheiro and Bates (2000), however, discuss simple plug-in predictors. These have been implemented by the authors in one of the “nlme” R software package. Predictors of this class can also be used for probability estimation, as considered by Hall and Clutter (2004).

Esteban et al. (2020) distinguish among the class of EPP estimates composite predictors based, inter alia, on area-level compositional models – a transformation of the multivariate Fay-Herriot model. These models are used in the analysis of compositional data, i.e. data that are a quantitative description of parts of some whole, conveying relative information about the variable under study. They can be expressed as proportions, percentages or probabilities. In labour force analyses, the indicators under study are sums or proportions of categories of a classification variable. They are therefore compositional parameters for domains that add up to one or to a known integer. This problem is discussed more extensively in the works of Aitchison (1986), edited by Pawlowsky-Glahn and Buccianti (2011). The authors proposed the use of predictors obtained by applying an alogist transformation to predict proportions (fractions). This transformation is often used for compositional data and models.

When discussing the issue of prediction using plug-in class predictors, the problem of calculating the MSE of these predictors should also be addressed. The analytical estimation of the MSE of the EP predictor has been addressed in the works of González-Manteiga et al. (2007), and Saei and Chambers (2003). In the remainder of this paper, we will consider MSE estimators of plug-in predictors obtained using a parametric bootstrap method.

4.3. EBP and plug-in predictors, and linear mixed models with correlated random effects vectors

This subsection is focused on the proposed use of EB and plug-in predictors under the assumption of a linear mixed model with correlated random effects vectors, which is given by the formula (1.131). In this case, we will denote the Best Predictor (BP) (4.2) by:

$$\hat{\theta}^{\rho}_{BP} = E(\theta|\mathbf{Y_s}). \tag{4.5}$$

It can therefore be determined analogously to the case where we assume a linear mixed model with uncorrelated random effects, as the conditional expectation value of the function of the random variables θ, assuming that the form of the conditional distribution $\mathbf{Y_r}|\mathbf{Y_s}$ is known. It should be noted that also in this case, the distribution depends in practice on the vector of unknown parameters $\boldsymbol{\tau}^{\rho}$. However, in this case, the auxiliary parameter ρ – the correlation coefficient between the random effects – needs to be included in this vector. By analogy with the EBP for models with uncorrelated random effects, if the vector $\boldsymbol{\tau}^{\rho}$ is replaced by its estimator, we obtain the empirical best predictor $\hat{\theta}^{\rho}_{EBP}$.

The plug-in predictor, assuming a linear model with correlated random effects vectors, will be denoted as follows:

$$\hat{\theta}^{\rho}_{PLUG-IN} = \theta\left(T^{-1}\left(\begin{bmatrix}\mathbf{Y}_s^T & \hat{\mathbf{Y}}^T_{r(\rho)}\end{bmatrix}^T\right)\right), \tag{4.6}$$

where $\hat{\mathbf{Y}}^T_{r(\rho)}$ is the vector of values obtained based on the model with correlated random effects vectors, which was assumed for the unobserved variables. It should be added that the use of the above predictors allows prediction over a wider range, it makes it possible to take into account the presence of correlations between random effects.

The problem of prediction using a predictor given by the formula (4.5) was considered in the paper by Krzciuk (2019). In the conducted simulation study, the author used data on the number of newly registered entities in the REGON register in 2017 in municipalities of south-western Poland. The size of the population in 2016 was used as an auxiliary variable, the voivodeship and type of municipality (urban, urban-rural and rural municipalities) were used as grouping variables. The simulation studies carried out were divided into three parts. The first two were based on a model-based approach. In the first part, data were generated based on a model with correlated, and in the second, with uncorrelated random effects. This made it possible to investigate the impact of model misspecification on the properties of the resulting estimates. The last part, on the other hand, was based on a design-based approach. In each part of the study, two predictors of the total value in the domain were considered – the best predictor and the empirical best predictor for each of the two models mentioned. It should be added that, in each part, the relative bias and mean squared error of all predictors considered in the analyses were determined by simulation. The results obtained in the simulation study suggest good properties of the proposed EB predictor accounting for random effects correlation, even in the case of the model misspecification considered in the second part – the average loss of accuracy was no more than 2%. In contrast, the first part of the study indicated the need to take into account the large number of random effects realisations and the sample size, and to estimate the correlation coefficient ρ with greater accuracy. On average, the decrease in accuracy resulting from the estimation of the model parameters, however, did not exceed 20%. For the design-based approach, both empirical versions of the considered predictors showed similar properties.

4.4. Estimation of the mean squared error of EBP and plug-in predictors

Following Diallo and Rao (2018), due to the lack of an explicit form of EBP mean squared error estimators, their second-order analytical approximations have been widely presented in the literature by Rao and Molina (2015), among others. It should be added, however, that it is not possible to use such approximations for complex non-linear parameters in the domain, even when the assumption regarding the normality of the prediction error distribution is met. Therefore, general-purpose methods, including bootstrap, are used for MSE estimation in this case. Estimators based on the parametric bootstrap method are analoguous to the EBLUP mean squared error estimators using the bootstrap model, which is given by the formula (3.38). One of the estimators belonging to this class is the estimator proposed by González-Manteiga et al. (2008). In this case, it will therefore have the following form:

$$M\hat{S}E^{boot}(\hat{\theta}_{\rho}^{EBP}) = B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_{\rho}^{EBP}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}^{*(b)}), \hat{\boldsymbol{\delta}}^{*(b)}) - \theta^{*(b)}\right)^{2}, \tag{4.7}$$

where, as in the case of the estimator given by formula (3.39), $\hat{\boldsymbol{\delta}}^{*(b)}$ is given by the same formula as $\hat{\boldsymbol{\delta}}$, where $\mathbf{Y}$ is replaced by $\mathbf{Y}^{*}$. Nevertheless, $\hat{\boldsymbol{\beta}}$ is the estimator obtained by the REML method and $\theta^{*(b)}$ is the value of the characteristic of interest θ obtained in the b-th implementation of the model (3.38) and $E^{*}(.)$ is the expected value in the bootstrap distribution.

Molina and Rao (2010) also note the possibility of using the double bootstrap method proposed in Hall and Maiti (2006) and described in the form of algorithm 1 where the MSE estimators at successive steps of the procedure are given by formulae (3.42) and (3.43). Using this method may provide better properties of the MSE estimator given the relative bias. However, the authors emphasise the importance of population size, as the method can be very time--consuming for large populations.

It should also be noted that the parametric bootstrap method can also be used to determine the mean squared error estimate of plug-in predictors. The estimator of the MSE discussed above considered by González-Manteiga et al. (2008) for this class of predictors can be written in the following form:

$$M\hat{S}E^{boot}(\hat{\theta}_{\rho}^{PLUG-IN}) = B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_{\rho}^{PLUG-IN}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}^{*(b)}), \hat{\boldsymbol{\delta}}^{*(b)}) - \theta^{*(b)}\right)^{2}, \tag{4.8}$$

where the notation is the same as in (4.7).

4.5. Applications of EB and plug-in predictors

In this subsection, selected areas of application of EBP and plug-in predictors in small area estimation will be presented, together with examples of an economic nature. The above predictors have been used in analyses concerning, among others, quality of life, unemployment, poverty, corporate finance, agricultural economics, and environmental economics.

The application of EBP in quality of life analyses was addressed by Chen and Liu (2019). The issue considered in the paper is the prediction of quantiles. In the analyses, the authors included, among others, the EBP considered by Molina and Rao (2010). The research conducted was based on data from the Survey of Labour and Income Dynamics conducted by Statistics Canada (2014). Henry et al. (2009) addressed the problem of predicting the value of total income for small areas. The analyses were conducted using data extracted by tax offices from tax returns in the United States. Molina and Martin (2018) also addressed the problem of predicting per capita income and average income in municipalities. The data included in the analyses came from the Mexican National Survey on Income and Expense of Households and the census. In addition, the authors compared the estimates obtained using EBP and selected other predictors.

Prediction of poverty measures using EBP was considered, among others, in the work of Krzciuk, Stachurski and Żądło (2017). The authors used data from Statistics Poland's survey of household budgets in their analysis. Among the characteristics of interest were the poverty rate and the poverty gap indicator. It should be added that the authors proposed a certain modification of the EBP. The properties of the aforementioned predictor were compared with the original EB-type predictor and the Hàjek estimator. Molina and Rao (2010), however, addressed the problem of predicting poverty rates as non-linear population parameters, using EBP. It should be added that the authors used data from the EU–SILC survey in the example analysed and the simulation studies carried out. The problem of predicting poverty rates using the EBP was also addressed by Boubeta et al. (2016). The authors made a comparison between this method and plug-in predictors. The issue of using the EBP to predict the percentage of people below the poverty line was also addressed by Boubeta et al. (2017). The authors used data from the Spanish Living Conditions Survey in their analysis. It should be added that Boubeta et al. (2017) included plug-in predictors in addition to the EBP in their study. The work of Elbers and van der Weide (2014) also addressed the problem of poverty prediction in the context of using the EBP and some modification of it. The authors used data from the Integrated Public Use Microdata Series (IPUMS) of the US population microcensus. The problem of prediction using plug-in predictors and the EBP in quality of life and unemployment issues was also considered by Hobza and Morales (2016). In their analyses, the authors used data from the Spanish Living Conditions Survey (SLCS) and the Labour Force Survey (SLFS) for Valencia. The parameter predicted by the authors was the unemployment rate. The possibility of using plug-in predictors in predicting the unemployment rate or the percentage of people with a certain economic activity status was also pointed out by Esteban et al. (2020). In their considerations, the authors used data from a variety of sources, including the Quality of Life Survey and unemployment registers.

The characteristics predicted by Esteban et al. (2020) included the proportions of people with a specific labour force participation status.

Żądło (2017), however, addresses the issue of predicting population and sub-population characteristics for future periods. In the simulation study carried out, the author used actual data on businesses in Polish counties derived from the Local Data Bank of Statistics Poland. The analyses considered, among other things, the EBP for parameters such as total value, median, standard and quarter deviation, and the classic asymmetry coefficient, as well as the EBLUP for total value.

Agricultural economics issues were addressed in the paper by Berg and Chandra (2014). The authors considered EBP and some modifications of it, as well as the direct predictor and the predictor considered by Karlberg (2000). The simulation study used artificial data analysed by Fuller (1991). Chandra et al. (2012a) considered two issues in their study in terms of prediction using EBPs of an economic nature. The first concerned the prediction of the percentage of farms with zero debt. The data included in these analyses came from the Australian Agricultural and Grazing Industries Survey. The second dataset considered was from the Albanian Living Standards Measurement Study. For this dataset, predictions were made on the proportion of households at risk of poverty and therefore with below-median incomes.

EBP and plug-in predictors have also found application in many other areas. In their paper Li and Lahiri (2007) considered predictors using the EBP to predict on-farm livestock numbers based on data from the Australian Agricultural and Grazing Industries Survey. Among the predictors considered by Li and Lahiri (2007) were, in addition to EBP, a predictor incorporating a logarithmic transformation, and an approximated BP and EBP. Salvati et al. (2012a) addressed the problem of predicting the proportion of lakes with low acid neutralisation rates. The analyses were based on data collected by the Environmental Monitoring and Assessment Program and concerned lakes in the northeastern US states.

4.6. Summary

This chapter focused on two classes of predictors – empirical best predictors and plug-in predictors. Section 4.1 presented the theoretical basis of prediction using the EBP, including a discussion of Monte Carlo approximation of predictors in this class.

The next subsection 4.2 dealt with plug-in predictors. In this section, the concept of a plug-in predictor was discussed along with a classification of predictors belonging to this class. The possibility of their application was also mentioned in theoretical terms.

Subsection 4.3 proposed the use of EBP and plug-in predictors under the assumption of linear mixed models with correlated random effects. An example of the application of the above predictor in economic research is also discussed.

In subsection 4.4, the estimation of the mean squared error of the prediction is addressed. Selected estimators for both EB and plug-in class predictors were presented.

The last subsection 4.5 showed selected applications of EB and plug-in predictors in studies of an economic nature. The areas of use of these predictors that were presented in this book are analyses of quality of life, poverty, unemployment, corporate finance, agricultural economics, and environmental economics.

The author's proposals presented in this chapter concern the use of EBP and plug-in predictors in prediction based on models with correlated random effects vectors and the estimator of MSEs for the classes of considered predictors.

Chapter 5

Simulation studies

This chapter presents the assumptions and results of the simulation studies carried out. It is also supplemented by a description of the considered dataset. The chapter will conclude with a summary of the obtained results. The main aim of the conducted analyses is a simulation comparison of the properties of the author's proposed predictors of some characteristics in domains, discussed in previous chapters, with corresponding predictors that do not take into account correlations between the random effects vectors and the selected estimators.

5.1. Dataset

The study variable for the analyses is the revenue of municipalities in million PLN. The data comes from the Local Data Bank of Statistics Poland and covers a three-year period (2018–2020). Total revenues of municipalities, in accordance with the Act of 13 November 2003 on revenues of local government units, consist of own revenues, subsidies, general subvention and funds for subsidising tasks. Municipalities' own revenues include revenues from the following taxes: real estate tax, agricultural tax, forest tax, vehicle tax, personal income tax, paid in the form of a tax card, tax on inheritances and donations, tax on civil law transactions, and revenues from additional tax liability related to tax avoidance. In addition, receipts from the following fees are also included: stamp duty, market fee, local, spa and dog ownership fee, advertising fee, mining fee (in the part specified in the Act of 9 June 2011. Geological and Mining Law) and others constituting municipal revenue, paid pursuant to separate regulations. Municipal revenues also include: income obtained by municipal budgetary units and payments from municipal budgetary establishments, income from municipal assets, inheritances, bequests and donations to the municipality, income from fines and penalties specified in separate regulations, income obtained for the benefit of the state budget in connection with the performance of tasks in the field of government administration and other tasks assigned by acts, unless separate regulations provide otherwise, interest on loans granted by the municipality, unless otherwise provided for in separate regulations, interest on untimely payments of receivables constituting the municipality's revenue, interest on funds accumulated in the municipality's bank accounts, unless otherwise provided for in separate regulations, subsidies from the budgets of other local government units, and other revenue due to the municipality under separate regulations. The general subsidy for municipalities consists of a levelling and balancing part. In addition, there is also an educational and development part for districts and provinces. The auxiliary variable, however, is the total population in municipalities in thousands of people in 2017–2019.

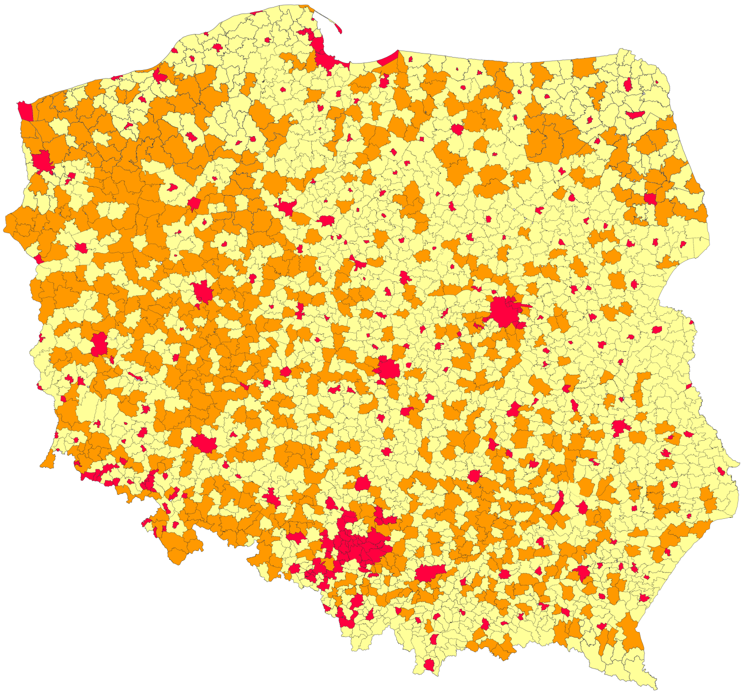


Figure 5.1. Map of municipalities in Poland

Note: Urban municipalities in red, urban-rural municipalities in orange and rural municipalities in yellow.

Source: Own elaboration.

The size of the population under consideration is $N = 7,398$ observations for three periods, and the sample size was set at approximately 20% of the population size ($n = 1,503$). Such a high ratio of the sample size to the population size was chosen to ensure high accuracy of the model parameter estimation, and the time-consuming nature of the calculations made it impossible to consider a population of a larger size. It should also be added that in one period, the sample size was 501 observations. In addition, the sample in the first period was drawn as a stratified sample, where the strata were defined on the basis of the affiliation of municipalities to voivodeships. The division of municipalities into domains was made on the basis of their belonging to 16 voivodeships and to two types of municipalities – rural and other (the number of domains was $D = 16 \times 2 = 32$). The whole sample, however, can be treated as a balanced panel according to the definition presented in subsection 2.3.1. Figure 5.1 shows the classification of the municipalities by type. Yellow colour represents rural municipalities, while the other colours correspond to urban (red) and urban-rural (orange) municipalities. Due to the time-consuming nature of the calculations in all the analyses presented in this book, only domains defined as rural municipalities belonging to particular voivodeships ($D = 16$) were considered in the prediction process. It is worth noting that the sample sizes of the domains were random. Figure 5.2 contains a map where the domains for rural municipalities are colour-coded. Each colour represents a different voivodeship. It should be added that voivodeships are, in accordance with the Regulation of the European Parliament and of the Council, level 2 units of the Classification of Territorial Units for Statistics (NUTS) in Poland, while municipalities are counted as local administrative units within this division.

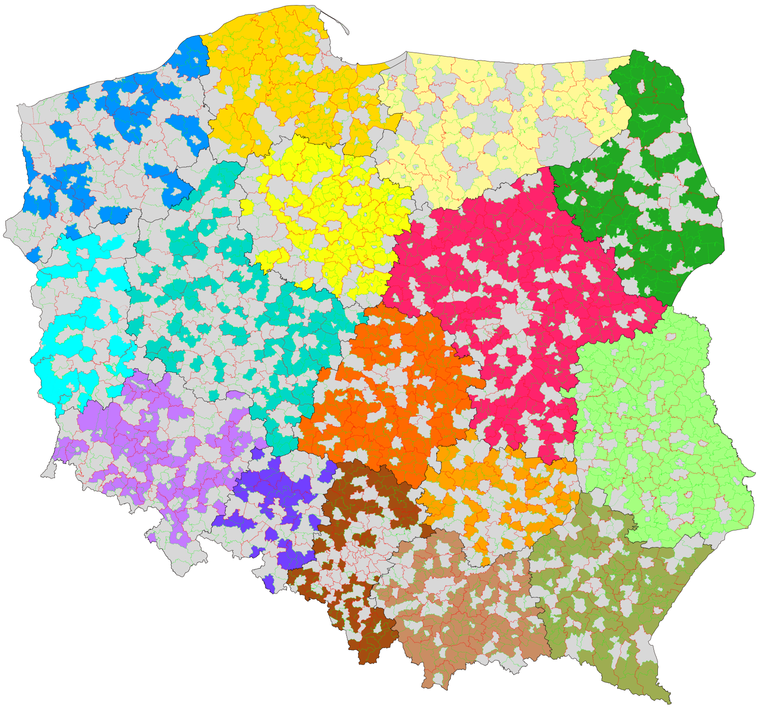


Figure 5.2. Domain map for rural municipalities

Source: Own elaboration

The analysis of the considered dataset began with determining selected descriptive statistics for the considered variables by domain. The median, mean value and coefficient of variation for the 16 domains considered in the prediction process were determined for the study and auxiliary variable, among others. It should be noted that in the case of the study variable, the coefficient of variation for each of the domains presented was significantly lower than for the population as a whole, for which it was 4.80. When only the domains of the rural municipalities were considered, it was 0.7 and therefore also higher than in more than two-thirds of the rural municipalities. This therefore suggests the use of a linear mixed model with domain-specific random effects.

5.2. Simulation study – variant I

The following subsection will present the assumptions and results obtained in the first of the simulation studies carried out. This as well as the following subsections will conclude with a brief summary of the most important results.

The study follows a model-based approach, hence the definitions used below such as bias, precision, accuracy or mean squared error will mean predictor bias, predictor precision, predictor accuracy and predictor mean squared error, respectively. The procedure for conducting the simulation study is presented in Algorithm 3.

Algorithm 3. Monte Carlo simulation study of the properties of selected predictors of the parameter θ in the domain

```
1. Estimation of the parameters of the model given by the formula (1.131) including a logarithmic transformation of the variables based on real population data using the REML method.
2. Drawing a sample according to the assumptions discussed in subsection 5.1.
3. Loop for $k = 1$ to $K$, where $K = 3000$:
   3.1. generation of population data of the study variable based on the actual values of the auxiliary variable and the estimated model parameters, including $\rho = -0.83$.
   3.2. determination of the values of the considered domain characteristics based on the generated data,
   3.3. determination of the values of the considered predictors and estimators of the domain characteristics based on the generated sample data.
   3.4. End of loop.
4. Determination of simulation values of selected accuracy and precision measurers.
```

It should be added that, in the case of the considered EB- and EBP-class predictors, the number of iterations $L = 300$ was assumed. For the considered statistics for the estimation of the selected domain characteristics, were determined, among others, simulations:

– relative bias of the predictor:

$$rB(\hat{\theta}_d) = \frac{\frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_d^k - \theta_d^k)}{\left|\frac{1}{K}\sum_{k=1}^{K}\theta_d^k\right|},$$

– relative mean prediction errors:

$$rD(\hat{\theta}_d - \theta_d) = rD(U_d) = \frac{\left(\frac{1}{K}\sum_{k=1}^{K}\left(U_d^k - \frac{1}{K}\sum_{j=1}^{K}U_d^j\right)^2\right)^{0.5}}{\left|\frac{1}{K}\sum_{k=1}^{K}\theta_d^k\right|},$$

– relative values of the root mean squared error of the prediction:

$$rRMSE(\hat{\theta}_d) = \frac{\left(\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\theta}_d^k - \theta_d^k\right)^2\right)^{0.5}}{\left|\frac{1}{K}\sum_{k=1}^{K}\theta_d^k\right|},$$

where K is the number of Monte Carlo iterations, $\hat{\theta}_d^k$ and θ_d^k are the value of the predictor and characteristics of the d-th domain in the k-th Monte Carlo iteration, respectively, and $U_d^k = = \hat{\theta}_d^k - \theta_d^k$. However, the results obtained for simulation of the relative bias and relative root mean squared error of prediction are discussed in more detail. The number of iterations adopted in the simulation study was considered sufficient – in the case of the overpopulation model with parameter values assumed to be at the level of estimates obtained from the full population data, relative values of simulation bias of the unbiased predictor *BP* as to modulus no greater than 0.3% were obtained for all cases considered. It should be added that the following simulation analyses presented include some modifications to Algorithm 3 and were preceded by preliminary studies with a reduced number of iterations. The results obtained indicated that in cases where the values of the parameter ρ were equal as to modulus, similar results were obtained. In the remainder of this chapter, only negative values of the ρ parameter consistent with the correlation between the random effects vectors for the considered actual population data are therefore considered. The the second variant of study considered the case of a stronger negative correlation ($\rho = -0.95$) and third a weaker negative correlation ($\rho = -0.65$), but the $\rho = -0.83$ from the first variant is obtained as population-based estimated parameter. In addition, in the case of the linear mixed model under consideration, a logarithmic transformation of the variables – the study and auxiliary variables – was included. Model selection was based on the Akaike (1973) information criterion (AIC). The significance of the parameters of the above model was verified using the tests presented in subsection 1.2.4, which are discussed in more detail in the works by Krzciuk and Żądło (2014a, 2014b), and Krzciuk (2018). Figure 5.3 presents a graph for the analysed variables, where a single broken line represents data from three periods for one municipality.

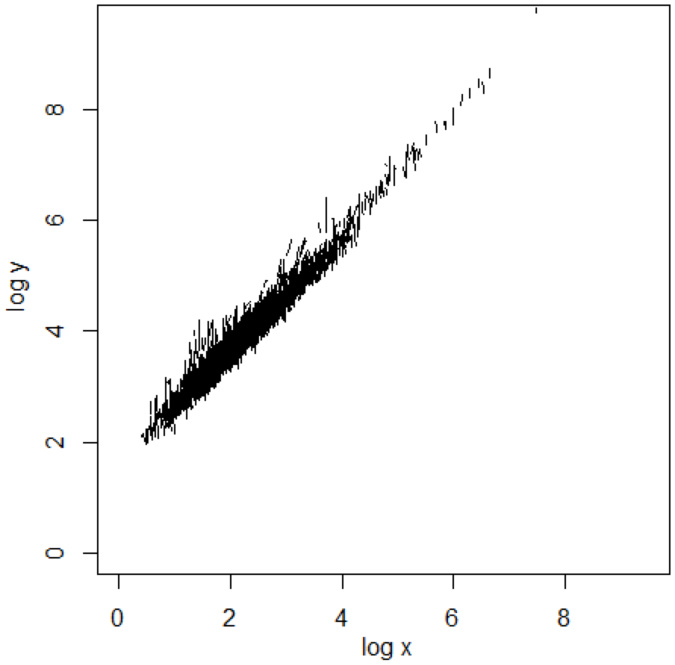


Figure 5.3. Scatter plot between $\log y$ and $\log x$

Source: Own elaboration.

This, as well as subsequent analyses, included the estimation of two characteristics – the total value and the median income of municipalities for all $D = 16$ domains. The following predictors were considered for each:

- proposed best predictor, taking into account the correlation of the random effects BP_ρ given by the formula (4.5) (denoted by BP_rho),
- proposed empirical best predictor, assuming a linear mixed model with correlated random effects EBP_ρ (3.33) (denoted by EBP_rho),
- empirical best predictor, assuming a model with uncorrelated random effects EBP_0 (denoted by EBP_0),
- proposed plug-in predictor, based on a linear mixed model with correlated random effects $PLUGIN_\rho$ given by the formula (4.6) (denoted by PLUGIN_rho),
- plug-in predictor, based on a linear mixed model with uncorrelated random effects $PLUGIN_0$ (denoted by PLUGIN_0).

The Horvitz-Thompson estimator (1.7) for the stratified sample (denoted by HT), the synthetic quotient estimator based on the HT estimator (Bracha 1996, p. 36) (SYNT) and the calibrated estimator (1.66) (CALIB) were also used to estimate total values in the domain. Median domain estimates, however, were also determined using the direct estimator considered by Särndal et al. (1992, p. 200) (denoted by SARNDAL) and the synthetic quotient estimator of the domain median proposed by Stachurski (2018, p. 511) (denoted by SYNT).

The plots and tables obtained in the first of the simulation studies carried out are presented below. Figure 5.4 shows box plots of the relative bias values $rB(.)$ as percentages for the considered predictors and estimators of total values in the domains. Each of the plots presents 16 simulation relative bias values as percentages. It can be seen that for the BP, EBP and plug-in predictors, the values of the above modulus are close to 0. They are clearly lower than the values of the relative bias moduli for the HT and calibrated estimator, for which it takes values of the order of several tens of percent.

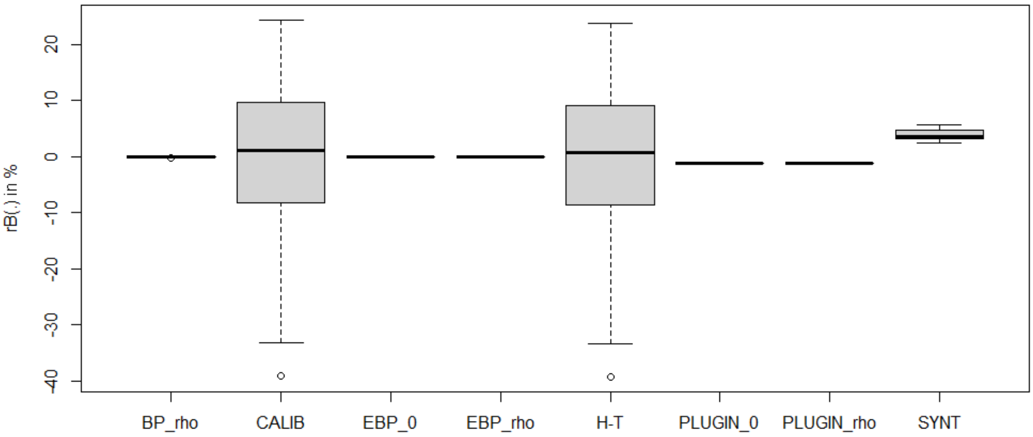


Figure 5.4. Values of $rB(.)$ of predictors and estimators of domain totals

Source: Own elaboration.

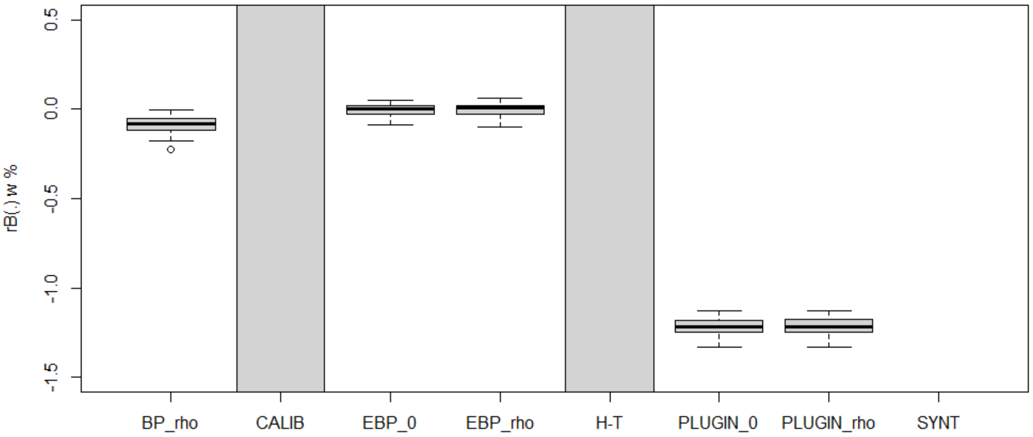


Figure 5.5. Selected $rB(.)$ values of predictors and estimators of domain totals

Source: Own elaboration.

Figure 5.5 presents an excerpt from Figure 5.4 allowing a more detailed analysis of the results for the BP, EBP and plug-in class predictors. It should be noted that the relative bias of the EBP_{ρ} predictor as far as the module is concerned did not exceed a value of 0.2%. In the case of the predictor $PLUGIN_\rho$, it is a value of 1.4%. It should be added, however, that only negative values of relative bias were obtained for this predictor.

Figure 5.6 presents box plots of the relative values of the root mean squared error $rRMSE(.)$ as percentages for the predictors and estimators of total values in the domains considered in the analysis. For all predictor proposals accounting for random effects correlation, the measure did not exceed a value of 3.3%, and for at least 75% of the domains, the value of 2.6% was not exceeded, as shown in Figure 5.7. In addition, the lowest median value of $rRMSE(.)$ was obtained for BP_ρ. This measure for at least 50% of the domains did not exceed 2%. The results obtained are therefore clearly lower than for the calibrated, HT or synthetic estimator, for which the median was over 5% and even in some cases over 9%.

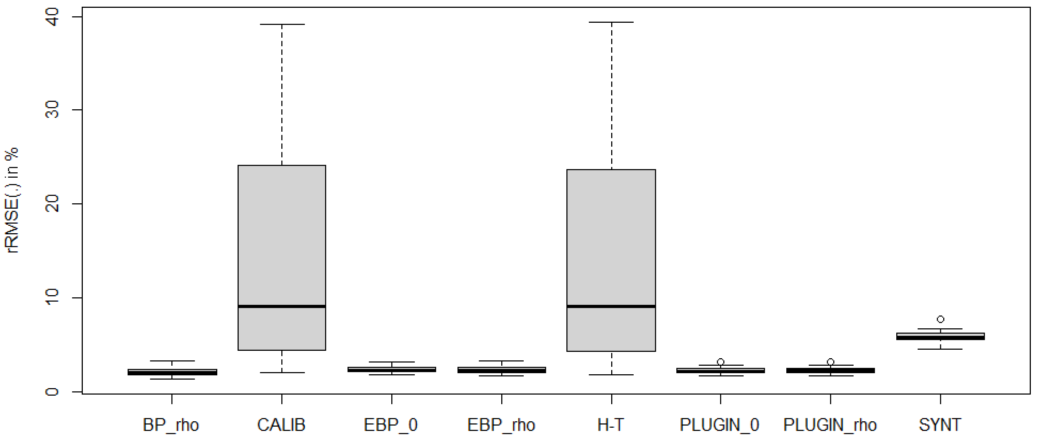


Figure 5.6. Values of $rRMSE(.)$ of predictors and estimators of domain totals

Source: Own elaboration.

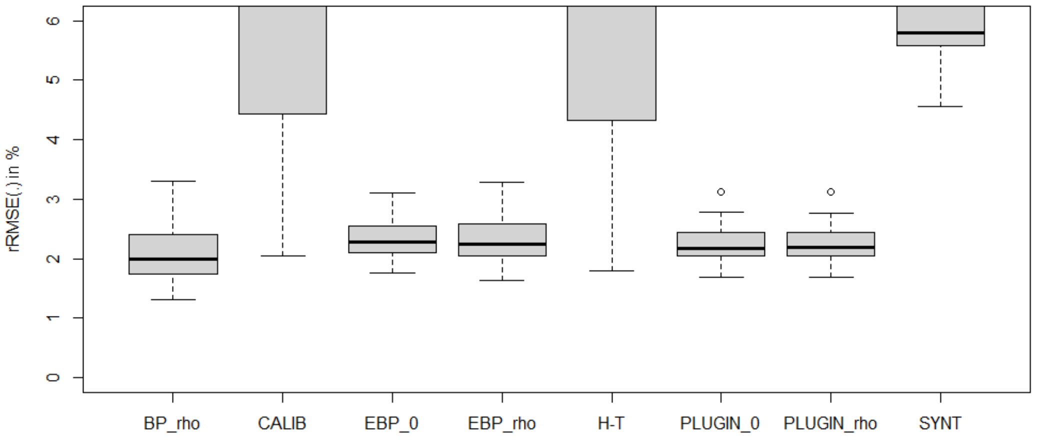


Figure 5.7. Selected $rRMSE(.)$ values of predictors and estimators of domain totals

Source: Own elaboration.

As for the relative values of the prediction standard error $rD(.)$, the lowest median value of this measure was obtained for the proposed predictors, i.e. BP_ρ and $PLUGIN_\rho$. In the case of the third, EBP_ρ, it was slightly higher, not exceeding 2.2%. In contrast, the highest $rD(.)$ results were obtained for the synthetic estimator. For the HT and calibrated estimator, as in the other cases, a significantly higher interquartile range was observed.

Table 5.1 presents the mean and median values of the ratio of accuracy measures of the BP_ρ predictor and the considered predictors and estimators of domain totals. The average increase in estimation accuracy resulting from the use of BP_ρ relative to EBP_0 was approximately 12%. Furthermore, for at least 50% of the domains, this gain is also 12%. In contrast, comparing BP_ρ with $PLUGIN_0$ and $PLUGIN_\rho$, the average gain in accuracy is approximately 9%. Note also that when comparing the proposed EBP-class predictor with the HT and calibrated estimators, the median gain in accuracy is as high as 75%. In the case of estimation precision, the average increase resulting from the use of BP_ρ over EBP_0 was approximately 12%, while the largest increase was observed over the synthetic estimator, over 50%.

Table 5.1. Mean and median values of the ratio of accuracy measures of BP_ρ and the considered predictors and estimators of domain totals

	$\frac{MSE(BP_\rho)}{MSE(.)}$		$\frac{RMSE(BP_\rho)}{RMSE(.)}$		$\frac{rRMSE(BP_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_ρ	0.78	0.74	0.88	0.86	0.88	0.86
EBP_0	0.78	0.77	0.88	0.88	0.88	0.88
$PLUGIN_\rho$	0.83	0.82	0.91	0.91	0.91	0.91
$PLUGIN_0$	0.84	0.83	0.91	0.91	0.91	0.91
HT	0.20	0.06	0.33	0.24	0.33	0.24
SYN	0.13	0.13	0.36	0.35	0.36	0.35
$CALIB$	0.17	0.06	0.31	0.25	0.31	0.25

Source: Own elaboration.

Table 5.2 shows the mean and median values of the ratio of accuracy measures of another of the predictor proposals that takes into account the correlation of random effects, the empirical version of BP_ρ (i.e. EBP_ρ) and the considered predictors and estimators of domain totals. When compared with EBP_0 for at least 50% of the domains, a gain in both accuracy and precision is observed. However, the largest average gain for both $rRMSE(.)$ and $D(.)$ is observed relative to the synthetic estimator at the level of several tens of percent. For the HT and calibrated estimator, however, a significant gain in precision of close to 60% is evident. For the precision of the estimates, similar results were obtained.

Table 5.2. Mean and median values of the ratio of accuracy measures of EBP_ρ and the considered predictors and estimators of domain totals

	$\frac{MSE(EBP_\rho)}{MSE(.)}$		$\frac{RMSE(EBP_\rho)}{RMSE(.)}$		$\frac{rRMSE(EBP_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_0	0.99	0.98	0.99	0.99	0.99	0.99
BP_ρ	1.30	1.35	1.14	1.16	1.14	1.16
$PLUGIN_\rho$	1.07	1.05	1.03	1.03	1.03	1.03
$PLUGIN_0$	1.07	1.06	1.03	1.03	1.03	1.03
HT	0.28	0.08	0.39	0.27	0.39	0.27
SYN	0.17	0.17	0.40	0.41	0.40	0.41
$CALIB$	0.23	0.08	0.36	0.29	0.36	0.29

Source: Own elaboration.

Table 5.3 presents the mean and median values of the ratio of accuracy measures of proposed $PLUGIN_\rho$ and the considered predictors and estimators of the domain totals. In the case of comparison with the predictor EBP_0, the average gain in accuracy is 4%. The largest gain of approximately 60% is seen for the synthetic estimator. A comparable gain of the accuracy also applies to the HT and the calibrated estimator. The gain of the precision when comparing with the EBP_0 predictor was higher, at 20%, and when comparing with the estimators, about a dozen percent.

Table 5.3. Mean and median values of the ratio of accuracy measures of $PLUGIN_\rho$ and the considered predictors and estimators of domain totals

	$\frac{MSE(PLUGIN_\rho)}{MSE(.)}$		$\frac{RMSE(PLUGIN_\rho)}{RMSE(.)}$		$\frac{rRMSE(PLUGIN_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_ρ	0.94	0.95	0.97	0.98	0.97	0.98
EBP_0	0.93	0.92	0.96	0.96	0.96	0.96
BP_ρ	1.23	1.22	1.11	1.10	1.10	1.10
$PLUGIN_0$	1.00	1.00	1.00	1.00	1.00	1.00
HT	0.26	0.07	0.38	0.26	0.38	0.26
SYN	0.15	0.15	0.39	0.39	0.39	0.39
$CALIB$	0.22	0.08	0.35	0.28	0.35	0.28

Source: Own elaboration.

The second parameter considered in the simulation study is the median in domain. Figure 5.8 presents box plots of the relative bias values of the considered predictors and estimators of the above parameter. For the BP- and EBP-class predictors, the lowest modulus results were obtained close to 0. The highest values were obtained for the estimator presented by Särndal et al. (1992, p. 200). As for the modulus, values in the order of several percent were obtained. In the case of the other statistics for estimating medians in the domain – plug-in predictors and the synthetic estimator – the results obtained as to the modulus did not exceed 10%. Figure 5.9 shows an excerpt from a plot of $rB(.)$ values allowing for a more detailed evaluation of the results obtained for the BP_ρ and EBP_ρ predictors. It should be noted that for both of these predictors as far as the module is concerned, the relative bias did not exceed the value of 0.3%.

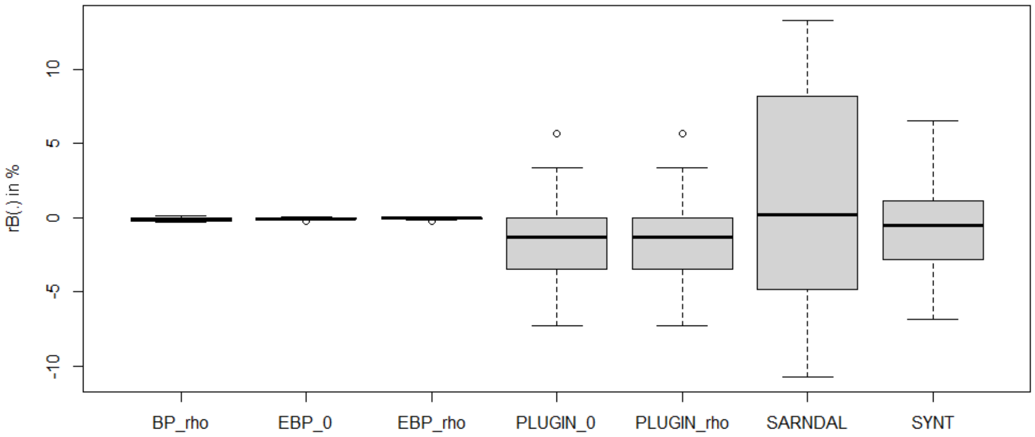


Figure 5.8. Values of $rB(.)$ of predictors and estimators of median in domain

Source: Own elaboration.

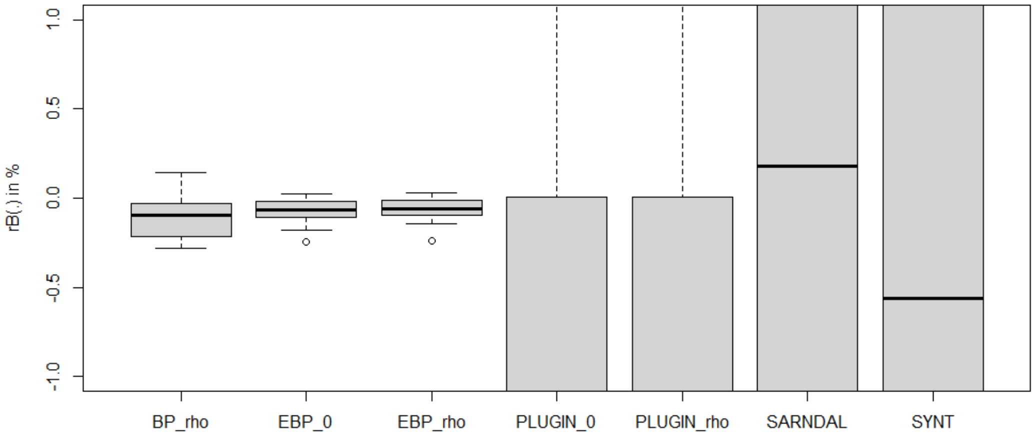


Figure 5.9. Selected $rB(.)$ values of predictors and estimators of median in domain

Source: Own elaboration.

Figure 5.10 shows the relative values of the root mean squared error $rRMSE(.)$. For the proposed EBP and BP predictors, the value of this measure did not exceed 4.5%. For the third proposed predictor accounting for correlated random effects, $PLUGIN_\rho$, this value was not exceeded for the median. For the two estimators considered in this study, the synthetic one and the one proposed by Särndal et al. (1992), the median relative $RMSE(.)$ was 5% and 9.5%, respectively.

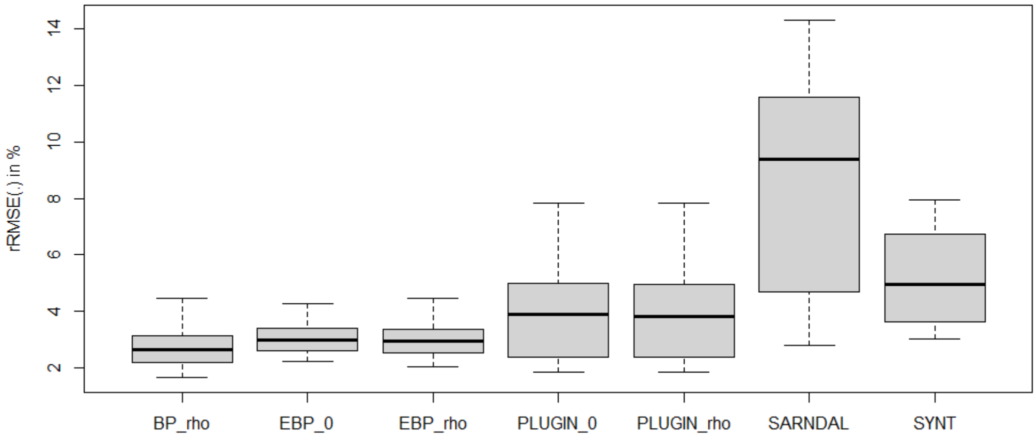


Figure 5.10. Values of $rRMSE(.)$ of predictors and estimators of median in domain

Source: Own elaboration.

In the case of $rD(.)$, the lowest median values for this measure were obtained for two of the proposed predictors: BP_ρ and $PLUGIN_\rho$. The highest values were obtained for the estimators included in the study – synthetic and Särndal et al. (1992), at 3.85% and 4.66%, respectively. The maximum value of $rD(.)$ did not exceed 5% for any of the considered predictors.

Table 5.4. Mean and median values of the ratio of accuracy measures of BP_ρ and the considered predictors and estimators of the median in domains

	$\frac{MSE(BP_\rho)}{MSE(.)}$		$\frac{RMSE(BP_\rho)}{RMSE(.)}$		$\frac{rRMSE(BP_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_ρ	0.80	0.80	0.89	0.89	0.89	0.89
EBP_0	0.78	0.79	0.88	0.89	0.88	0.89
$PLUGIN_\rho$	0.62	0.66	0.76	0.81	0.76	0.81
$PLUGIN_0$	0.62	0.65	0.76	0.81	0.76	0.81
$SARN$	0.17	0.11	0.38	0.34	0.38	0.34
SYN	0.33	0.31	0.56	0.55	0.56	0.55

Source: Own elaboration.

Table 5.4 presents the mean values and median of the ratio of accuracy measures of the predictor BP_ρ and the considered predictors and estimators of the median in domains. It should be noted that when comparing the accuracy of BP_ρ with EBP_0 and $PLUGIN_0$, an average gain of at least several percent is apparent, at 12% and 24%, respectively. When making a comparison with the considered estimators, however, this is a gain of several tens of percent. The average gain in precision, nevertheless, is as high as 60% when comparing against the considered estimators and 11% when analysed against EBP_0.

Table 5.5. Mean and median values of the ratio of accuracy measures of EBP_ρ and the considered predictors and estimators of the median in domains

	$\frac{MSE(EBP_\rho)}{MSE(.)}$		$\frac{RMSE(EBP_\rho)}{RMSE(.)}$		$\frac{rRMSE(EBP_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_0	0.97	0.97	0.98	0.98	0.98	0.98
BP_ρ	1.26	1.25	1.12	1.12	1.12	1.12
$PLUGIN_\rho$	0.81	0.90	0.86	0.94	0.86	0.94
$PLUGIN_0$	0.80	0.88	0.86	0.93	0.86	0.93
$SARN$	0.22	0.13	0.43	0.37	0.43	0.37
SYN	0.41	0.39	0.62	0.62	0.62	0.62

Source: Own elaboration.

Table 5.5 shows the mean values and the median of the ratio of the accuracy measures of EBP_ρ and the considered predictors and estimators of the median in domains. When comparing the accuracy of EBP_ρ and a predictor of this class that does not take into account random effects correlations, an average gain of 2% is evident. However, making a comparison between the accuracy measure and the results for $PLUGIN_0$, an average gain of 14% is observed. The largest gain in accuracy is seen when comparing the results for the proposed EBP-class predictor with the results obtained for the analysed median domain estimators. In this case, the average gain was even 57%. For the precision of the estimates, similar results were obtained.

The last table shows the mean and median values of the ratio of the accuracy measures of $PLUGIN_\rho$ and the considered the domains' median predictors and estimators. The largest gain in accuracy, but also in precision, was achieved when comparing $PLUGIN_\rho$ against the domain's median estimators considered in the study. Furthermore, in the case of precision, an average gain of 1% can be observed when comparing with the predictor $PLUGIN_0$ and 13% when comparing with the predictor EBP_0. When comparing with the synthetic estimator, the average gain is 43% on precision and 24% on accuracy, and with the estimator of Särndal et al. (1992) in both cases it is 43%.

Table 5.6. Mean and median values of ratio of accuracy measures of $PLUGIN_{\rho}$ and the considered predictors and estimators of the median in domains

	$\frac{MSE(PLUGIN_{\rho})}{MSE(.)}$		$\frac{RMSE(PLUGIN_{\rho})}{RMSE(.)}$		$\frac{rRMSE(PLUGIN_{\rho})}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_{ρ}	1.96	1.15	1.31	1.07	1.31	1.07
EBP_0	1.93	1.14	1.29	1.06	1.29	1.06
BP_{ρ}	2.39	1.51	1.45	1.23	1.45	1.23
$PLUGIN_0$	0.99	1.00	1.00	1.00	1.00	1.00
$SARN$	0.46	0.23	0.57	0.48	0.57	0.48
SYN	0.60	0.63	0.76	0.79	0.76	0.79

Source: Own elaboration.

In summary, relative bias values of no more than 0.3% were obtained for the proposed EBP- and BP-class predictors and 1.4% and 10% for the plug-in predictor when the total and median values in the domains were predicted, respectively. The simulation-derived values of relative $RMSE(.)$ these predictors were no higher than 3.3% when the characteristic of interest was the total value and 4.5% – when the median in domain. Comparing the properties of the predictor proposals with selected predictors that do not take correlated random effects into account, the average gain in accuracy and precision ranged from a few to several percent. Compiled with the considered estimators, gains of up to several tens of percent were noted. For selected domains, the gain in accuracy was as high as 92% and in precision 66%.

5.3. Simulation study – variant II

This subsection will present the assumptions and results obtained in the second simulation study carried out. It is a modification of the analyses discussed in the previous subsection. In this study, the values of the study variable are generated according to Algorithm 3, where $\rho = -0.95$ is assumed. Thus, the case of a stronger correlation between random effects than for the original population data but with a direction according to the original value of this parameter is considered. All other parameters for both fixed and random effects were assumed to be in accordance with the estimates for the population data.

Figure 5.11 shows the relative biases of the statistics considered. Similar to Figure 5.4, absolute values of $rB(.)$ close to 0 were obtained in variant one for all the proposed predictors of the total value in the domain considering random effects correlation. It should be noted that for EBP_{ρ}, values of relative bias moduli not exceeding 0.08% were obtained. This can be seen more precisely in Figure 5.12 showing selected values of relative bias of the considered statistics. For the considered estimators, again significantly higher values of $rB(.)$ reaching as far as 40% in modulus were obtained.

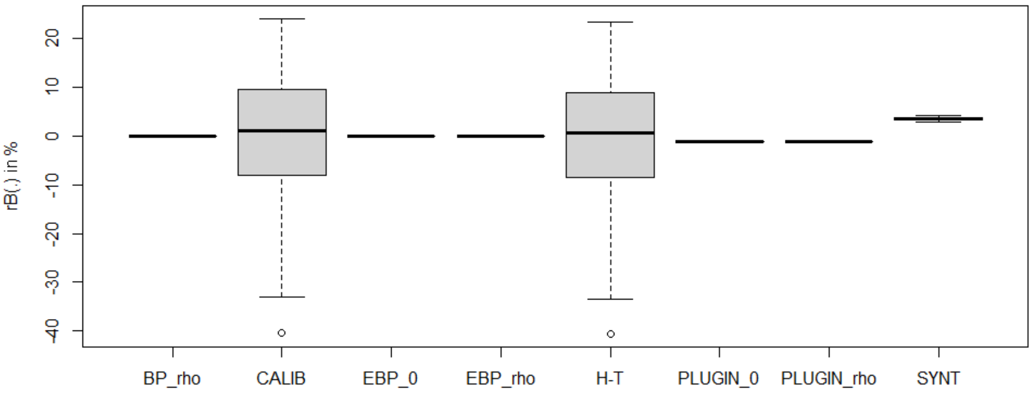


Figure 5.11. Values of $rB(.)$ of predictors and estimators of domain totals ($\rho = -0.95$)

Source: Own elaboration.

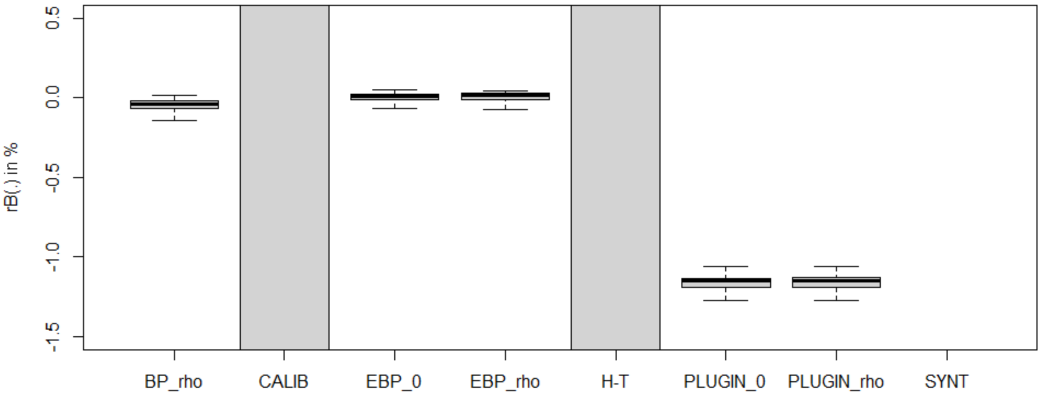


Figure 5.12. Selected $rB(.)$ of predictors and estimators of domain totals ($\rho = -0.95$)

Source: Own elaboration.

Figures 5.13 and 5.14 show the relative $RMSE(.)$ values of the predictors and estimators considered in the study. Comparing these with the plots presented in Figures 5.6 and 5.7, it can be seen that significantly better results were obtained for the proposed predictors. The maximum value of the relative root mean squared error of the prediction for none of them exceeded 2.9%. In subsection 5.2.2, the value was approximately 0.4 percentage points higher. In addition, for each of these predictors, the third quartile did not exceed a value of 2.25% (in the first study, it was 2.6%). Comparing also the median values for each of the domain total value estimates in this variant, the lowest result of just under 1.75% was obtained for BP_ρ. In contrast, the highest values were obtained for the HT and calibrated estimators. In their case, in at least 50% of the domains, the value of $rRMSE(.)$ was more than 9%, as in subsection 5.2.2.

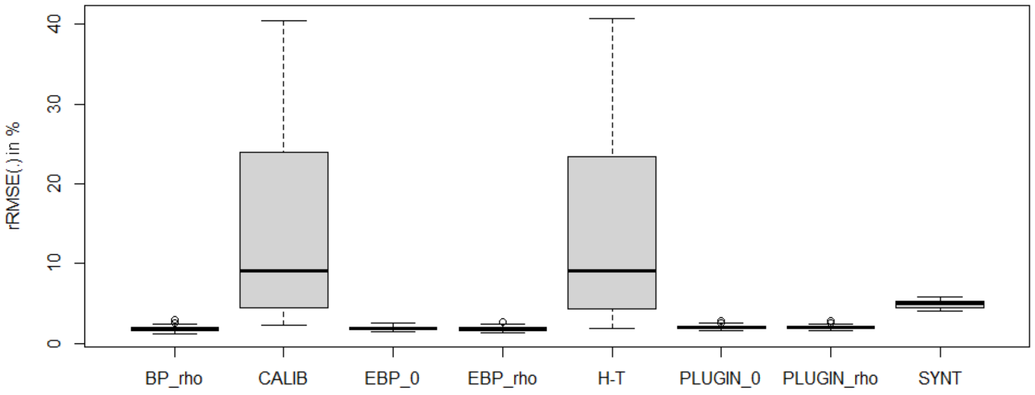


Figure 5.13. Values of $rRMSE(.)$ of predictors and estimators of domain totals ($\rho = -0.95$)

Source: Own elaboration.

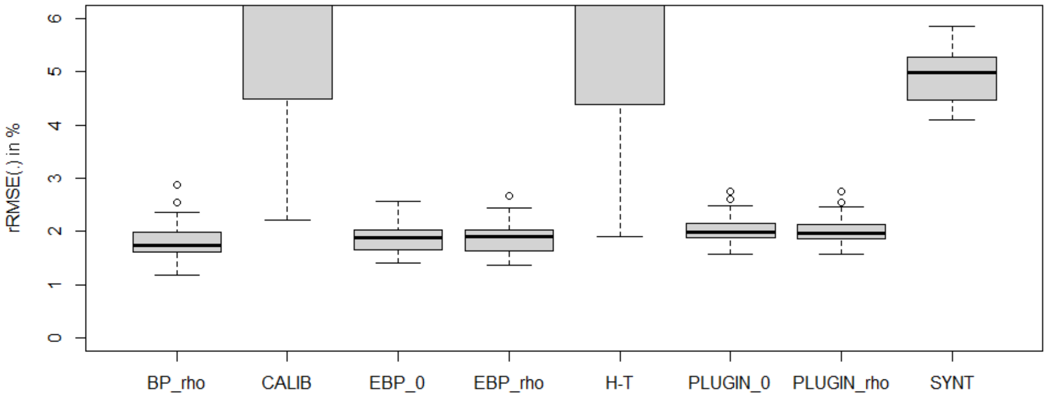


Figure 5.14. Selected $rRMSE(.)$ values of predictors and estimators of domain totals ($\rho = -0.95$)

Source: Own elaboration.

In the case of the relative values of the standard error of estimates considered in the paper for both predictors and estimators, it can be noted that for each of the analysed predictor proposals, taking into account correlated random effects, the median $rD(.)$ is lower than 2%. The highest relative values of the estimation standard error were obtained for the synthetic estimator, for which the minimum value was even more than 50% higher than the median value of the proposed predictors.

Table 5.7 presents the mean values and the median of the ratio of the accuracy measures of BP_ρ and the considered predictors and estimators of the domain totals. When comparing the accuracy of BP_ρ with the other predictors of the EBP and BP classes, an average gain in accuracy of approximately 3–5% can be observed. In contrast, compared to plug-in predictors, the average gain in accuracy can be close to 10%. The highest values for the average gain in accuracy were obtained when comparing with the HT and calibrated estimator – they were over 70%. It should also be added that the median gain in accuracy in this case is close to 80%. As in subsection 5.2.2, the largest average gain in precision was observed in relation to the synthetic estimator and is of the order of several tens of percent.

Table 5.7. Mean and median values of the ratio of accuracy measures of BP_ρ and the considered predictors and estimators of domain totals ($\rho = -0.95$)

	$\frac{MSE(BP_\rho)}{MSE(.)}$		$\frac{RMSE(BP_\rho)}{RMSE(.)}$		$\frac{rRMSE(BP_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_ρ	0.92	0.92	0.95	0.96	0.95	0.96
EBP_0	0.94	0.96	0.97	0.98	0.97	0.98
$PLUGIN_\rho$	0.81	0.81	0.90	0.90	0.90	0.90
$PLUGIN_0$	0.80	0.80	0.89	0.89	0.89	0.89
HT	0.12	0.05	0.27	0.22	0.27	0.22
SYN	0.15	0.16	0.38	0.40	0.38	0.40
$CALIB$	0.11	0.05	0.26	0.23	0.26	0.23

Source: Own elaboration.

Table 5.8. Mean and median values of the ratio of accuracy measures of EBP_ρ and the considered predictors and estimators of domain totals ($\rho = -0.95$)

	$\frac{MSE(EBP_\rho)}{MSE(.)}$		$\frac{RMSE(EBP_\rho)}{RMSE(.)}$		$\frac{rRMSE(EBP_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_0	1.02	1.02	1.01	1.01	1.01	1.01
BP_ρ	1.11	1.08	1.05	1.04	1.05	1.04
$PLUGIN_\rho$	0.88	0.90	0.94	0.95	0.94	0.95
$PLUGIN_0$	0.87	0.89	0.93	0.94	0.93	0.94
HT	0.14	0.05	0.29	0.23	0.29	0.23
SYN	0.16	0.15	0.39	0.39	0.39	0.39
$CALIB$	0.12	0.06	0.28	0.24	0.28	0.24

Source: Own elaboration.

Table 5.8 shows the results of the analysis of the mean and median values of the ratio of the accuracy measures of EBP_ρ and the analysed predictors and estimators of the domain totals.

Comparing the accuracy of the EBP-class predictor proposal, which takes correlated random effects into account, with plug-in predictors, it can be seen that for at least 50% of the domains, the gain in accuracy is about 5%. However, making a comparison with the estimators, including the HT estimator and the calibrated estimator, the average gain even reaches more than 70%. In contrast, the largest gain in precision was observed with the synthetic estimator, amounting to more than 40% for at least 50% of the domains.

Table 5.9 shows the mean and median values of the ratio of the accuracy measures of the plug-in predictor proposal $PLUGIN_\rho$ and the included predictors and estimators of the domain totals. When considering the accuracy of the $PLUGIN_\rho$ estimates and the estimators included in the study, it can be seen that the median gain in accuracy ranges from 57% for the synthetic estimator to as much as 76% for the HT estimator. In the case of precision, when comparing the predictor $PLUGIN_\rho$ with the other predictors included in the analysis, an average gain of up to several percent was noted. When it was compared with the analysed estimators, this gain reached up to 50%.

Table 5.9. Mean and median values of the ratio of accuracy measures of $PLUGIN_\rho$ and the considered predictors and estimators of the domain totals ($\rho = -0.95$)

	$\frac{MSE(PLUGIN_\rho)}{MSE(.)}$		$\frac{RMSE(PLUGIN_\rho)}{RMSE(.)}$		$\frac{rRMSE(PLUGIN_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_ρ	1.14	1.11	1.07	1.05	1.07	1.05
EBP_0	1.16	1.17	1.08	1.08	1.08	1.08
BP_ρ	1.27	1.24	1.12	1.11	1.12	1.11
$PLUGIN_0$	0.99	0.98	0.99	0.99	0.99	0.99
HT	0.17	0.06	0.31	0.24	0.31	0.24
SYN	0.18	0.19	0.42	0.43	0.42	0.43
$CALIB$	0.14	0.07	0.30	0.26	0.30	0.26

Source: Own elaboration.

In this variant of the simulation study, the problem of estimating the median in domains is also included. Figure 5.15 includes the results for the relative bias $rB(.)$. For all predictors and estimators except for plug-in predictors, the median value did not exceed 1% in modulus. For the predictors of the mentioned class, it was about 1.3%. However, it should be noted that for the estimators considered, the maximum value exceeded even more than a dozen percent. For the plug-in predictors, it was less than 7.5%, and for the proposed EBP- and BP-class predictors, it was 0.25%. The results for the latter two groups of predictors can be seen in more detail in Figure 5.16.

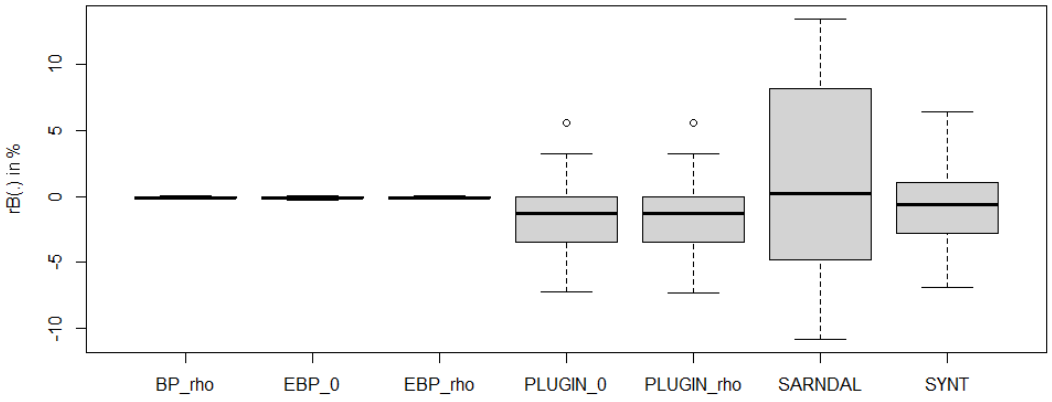


Figure 5.15. Values of $rB(.)$ of predictors and estimators of median in domains ($\rho = -0.95$)

Source: Own elaboration.

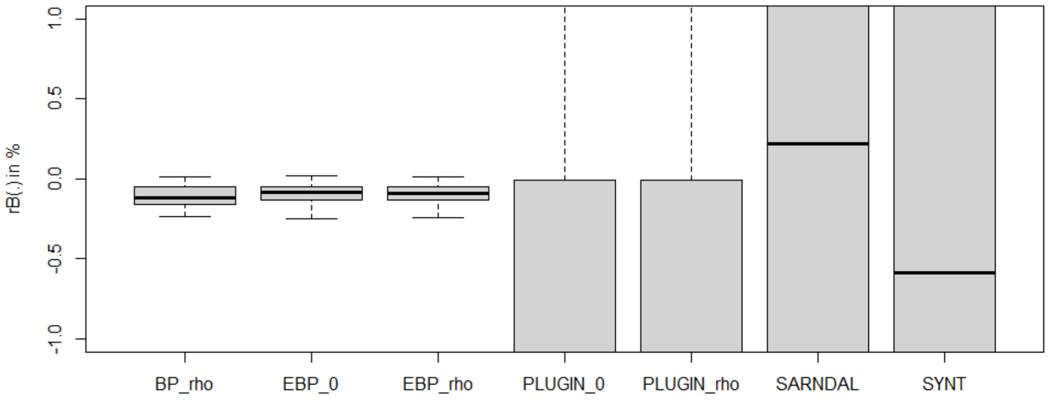


Figure 5.16. Selected $rB(.)$ values of predictors and estimators of median in domains ($\rho = -0.95$)

Source: Own elaboration.

Figure 5.17 shows plots of the relative values of the root $MSE(.)$ of the analysed predictors and estimators of the domain median. Comparing this and the first variant of the simulation study, it should be noted that the maximum value of this measure for EBP_ρ and BP_ρ in this variant did not exceed 4.1%, while in the previous variant, this value was 0.4 percentage points higher. For $PLUGIN_\rho$, the median value of $rRMSE(.)$ in this part of the analyses is lower, at around 3.7%. The highest results were again obtained for the estimator presented by Särndal et al. (1992, p. 200), for which $rRMSE(.)$ is even close to 14%.

In the case of $rD(.)$ for the analysed predictors and estimators of the considered parameter in the domains, comparing this variant of the simulation study with the results presented in subsection 5.2.2, it can be observed that for all predictors, the maximum value is lower and amounts to no more than 4.5%. Among the estimators, a value of $rD(.)$ higher than 5% was obtained for the estimator of Särndal et al. (1992, p. 200). It even amounted to more than 9%. Note that for BP_ρ and $PLUGIN_\rho$, the value of $rD(.)$ was no higher than 2.5% for 50% of the domains.

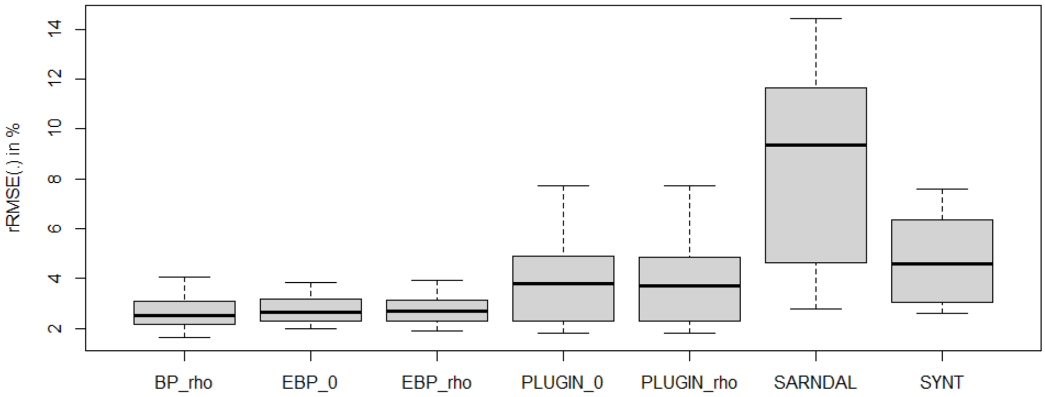


Figure 5.17. Values of $rRMSE(.)$ of predictors and estimators of median in domains ($\rho = -0.95$)

Source: Own elaboration.

Table 5.10 presents the mean and median values of the ratio of the accuracy measure of BP_ρ and the analysed predictors and estimators of the second parameter considered, i.e. the median in domain. By analogy with subsection 5.2.2, when comparing the values of the prediction accuracy measure for BP_ρ with $PLUGIN_0$, one should notice an average gain in accuracy of even more than 20%. Comparing with the estimators, it is again of the order of 40–60%. In the case of precision, however, the gain was approx. 5% when comparing with EBP-class predictors and even 56% with estimators.

Table 5.10. Mean and median values of the ratio of the accuracy measures of BP_ρ and the considered predictors and estimators of the median in domains ($\rho = -0.95$)

	$\frac{MSE(BP_\rho)}{MSE(.)}$		$\frac{RMSE(BP_\rho)}{RMSE(.)}$		$\frac{rRMSE(BP_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_ρ	0.91	0.93	0.95	0.97	0.95	0.97
EBP_0	0.90	0.92	0.95	0.96	0.95	0.96
$PLUGIN_\rho$	0.63	0.68	0.77	0.83	0.77	0.83
$PLUGIN_0$	0.62	0.66	0.76	0.81	0.76	0.81
$SARN$	0.17	0.11	0.37	0.33	0.37	0.33
SYN	0.39	0.34	0.61	0.58	0.61	0.58

Source: Own elaboration.

Table 5.11. Mean and median values of the ratio of the accuracy measures of EBP_ρ and the considered predictors and estimators of the median in domains ($\rho = -0.95$)

	$\frac{MSE(EBP_\rho)}{MSE(.)}$		$\frac{RMSE(EBP_\rho)}{RMSE(.)}$		$\frac{rRMSE(EBP_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_0	0.99	0.99	0.99	0.99	0.99	0.99
BP_ρ	1.12	1.07	1.06	1.03	1.06	1.03
$PLUGIN_\rho$	0.72	0.82	0.81	0.90	0.81	0.90
$PLUGIN_0$	0.70	0.80	0.80	0.89	0.80	0.89
$SARN$	0.20	0.11	0.40	0.33	0.40	0.33
SYN	0.44	0.44	0.64	0.66	0.64	0.66

Source: Own elaboration.

Table 5.11 presents the results of the accuracy analysis of EBP_ρ and the other predictors and estimators of the median in domains. When making a comparison of the $RMSE(.)$ predictors EBP_ρ and EBP_0 similarly to subsection 5.2.2, an average increase in accuracy is observed. When compared with the plug-in predictor, which does not take random effects correlation into account, the average increase in accuracy is 20%, which is approximately six percentage points higher than in the first variant of the analyses. The highest values of the average gain in accuracy are observed when comparing with the estimators presented in the study. It can be of the order of several tens of percent. Similar results were also obtained for the precision measures.

Table 5.12. Mean and median values of the ratio of the accuracy measures of $PLUGIN_\rho$ and the considered predictors and estimators of the median in domains ($\rho = -0.95$)

	$\frac{MSE(PLUGIN_\rho)}{MSE(.)}$		$\frac{RMSE(PLUGIN_\rho)}{RMSE(.)}$		$\frac{rRMSE(PLUGIN_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_ρ	2.26	1.25	1.39	1.11	1.39	1.11
EBP_0	2.26	1.26	1.39	1.12	1.39	1.12
BP_ρ	2.44	1.47	1.46	1.21	1.46	1.21
$PLUGIN_0$	0.98	0.99	0.99	1.00	0.99	1.00
$SARN$	0.45	0.22	0.56	0.47	0.56	0.47
SYN	0.69	0.78	0.82	0.88	0.82	0.88

Source: Own elaboration.

The last of the tables presented in this subsection, Table 5.12, shows the mean and median values of the ratio of accuracy measures of $PLUGIN_\rho$ and the selected predictors and estimators of the median in domains. Comparing $PLUGIN_\rho$ with a predictor of the same class, not taking random effects correlation into account, as in the first study, the average gain in accuracy is 1%. The highest values of the average gain in accuracy as well as precision are observed when juxtaposed with the estimators under consideration and can be as high as more than 40%. Furthermore, in the case of precision, the average gain, when compared with the EBP-class predictors, is around 7–8%.

In summary, similar to the analyses in subsection 5.2, relative simulation biases close to 0 were obtained for the proposed EBP and BP predictors. The simulation-derived $rRMSE(.)$ values of the proposed predictors were no higher than 3.0% when the characteristic of interest was the total value and 4.1% when the median in domain. It should be added that the above maximum values are lower than in the first study variant. When compared with the considered predictors, the average gain in accuracy and precision ranged from a few to even several percent. When comparing the properties of the proposed predictors with the estimators considered in the analyses, it was noted to be as high as several tens of percent. The maximum gain values were 94% and 65% on accuracy and precision, respectively.

5.4. Simulation study – variant III

The following subsection will present the assumptions and results obtained in the third simulation study carried out. Like the second variant, it introduces some changes to the analyses in subsection 5.2. This study was carried out according to Algorithm 3, presented in subsection 5.2.1, however, $\rho = -0.65$ was assumed. In this subsection, therefore, the case of a weaker correlation between random effects than for the original data but with a direction according to the original parameter value is analysed. It should be added that the values of the other model parameters were assumed to be in accordance with the estimates obtained for the population.

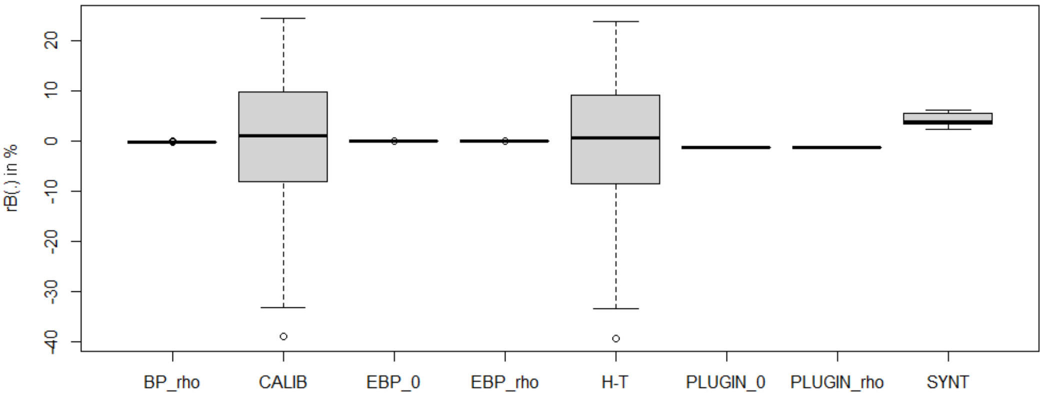


Figure 5.18. Values of $rB(.)$ of predictors and estimators of domain totals ($\rho = -0.65$)

Source: Own elaboration.

Figure 5.18 presents box plots of $rB(.)$ predictors and estimators of the total values in the domains. As with the other two analyses, relative bias values as to modulus close to 0 were obtained for all three proposed predictors. Analysing the plot in Figure 5.19, it can be seen that for the EBP_ρ predictor, the modulus did not exceed 0.1%. In the case of the $PLUGIN_\rho$, they were no higher than 1.43%. These values are therefore similar to the results obtained in subsection 5.2.2. The highest relative values of prediction bias were obtained for the calibrated and HT estimators, even on the order of several tens of percent.

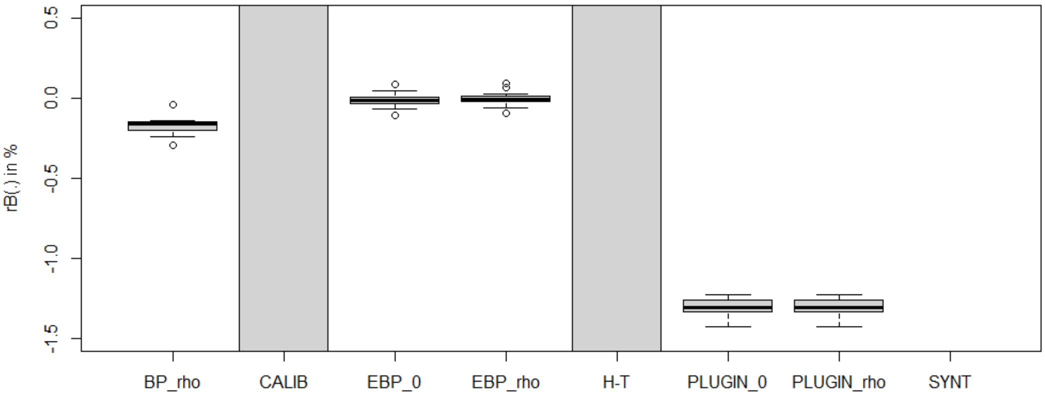


Figure 5.19. Selected $rB(.)$ values of predictors and estimators of domain totals ($\rho = -0.65$)

Source: Own elaboration.

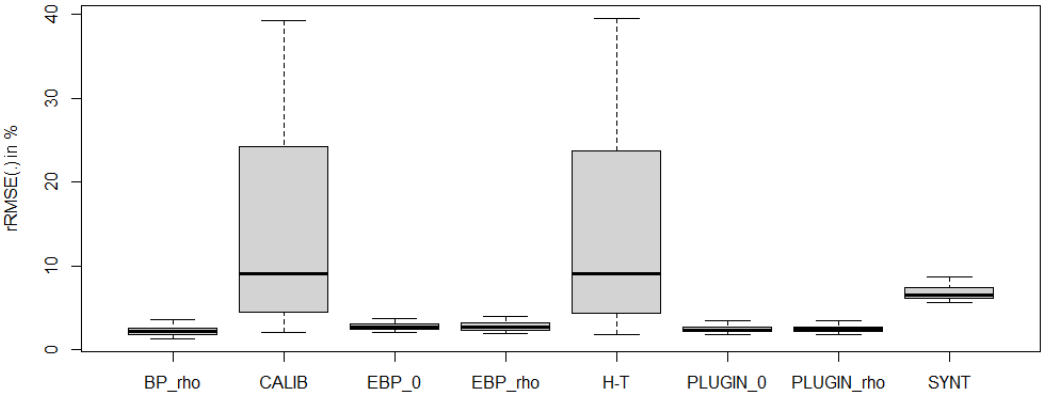


Figure 5.20. Values of $rRMSE(.)$ of predictors and estimators of domain totals ($\rho = -0.65$)

Source: Own elaboration.

Figures 5.20 and 5.21 show the results for $rRMSE(.)$ predictors and estimators of the total values in the domains in the form of box plots. For the proposed predictors that take into account the correlation between random effects, the value of 4% was not exceeded. Despite accounting for a weaker correlation in this study than for the original data, this value is only a few tenths of a percentage point higher than for the case considered in subsection 5.2.2. The median of $rRMSE(.)$ of these predictors, however, did not exceed 2.8%. For the considered estimators, i.e. synthetic, HT and calibrated, the median of $rRMSE(.)$ is 6%-9%. For the latter two estimators, $rRMSE(.)$ even reaches more than 39%, similar to the first variant of the study.

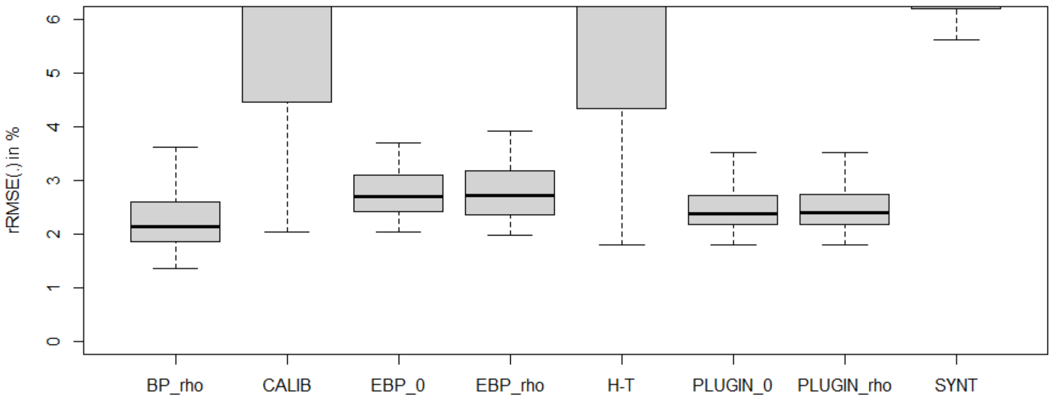


Figure 5.21. Selected $rRMSE(.)$ values of predictors and estimators of domain totals ($\rho = -0.65$)

Source: Own elaboration.

In the case of the analysis of the relative $D(.)$ values of the predictors and estimators of the total values in the domains, however, it is possible to observe that for all statistics, the $rD(.)$ values did not exceed 4%, similarly to the analyses presented in subsection 5.2.2. The exception is the synthetic estimator, for which the minimum value is about 4.1%. It should be added that in the case of BP_ρ and $PLUGIN_\rho$, the median relative values of the prediction standard error were no higher than 2.2%.

Table 5.13 presents the mean and median values of the ratio of the accuracy measures of BP_ρ and all considered predictors and estimators of the domain totals. When comparing BP_ρ with the analysed predictors of the EBP class, it should be noted that the average gain in accuracy is about 20%. In subsection 5.2.2, this value was about 8 percentage points lower. It is furthermore twice as high as the average gain in accuracy when comparing BP_ρ with plug-in predictors, a result close to that obtained in the first variant of the study. However, the highest values of the average gain in accuracy are seen in relation to the analysed estimators, amounting to more than 60%. For precision, similar results were obtained.

Table 5.13. Mean and median values of the ratio of accuracy measures of BP_{ρ} and the considered predictors and estimators of the domain totals ($\rho = -0.65$)

	$\frac{MSE(BP_{\rho})}{MSE(.)}$		$\frac{RMSE(BP_{\rho})}{RMSE(.)}$		$\frac{rRMSE(BP_{\rho})}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_{ρ}	0.63	0.61	0.79	0.78	0.79	0.78
EBP_0	0.64	0.60	0.80	0.77	0.80	0.77
$PLUGIN_{\rho}$	0.80	0.79	0.89	0.89	0.89	0.89
$PLUGIN_0$	0.81	0.80	0.90	0.90	0.90	0.90
HT	0.23	0.07	0.35	0.26	0.35	0.26
SYN	0.12	0.11	0.33	0.33	0.33	0.34
$CALIB$	0.19	0.08	0.33	0.27	0.33	0.27

Source: Own elaboration.

Table 5.14 shows the mean and median values of the ratio of the accuracy measures of EBP_{ρ} and the considered predictors and estimators of the domain totals. In the case of precision, the average gain relative to each of the estimators under consideration is similar, at more than 50%. In contrast, the largest average gain in precision was observed when comparing with the synthetic predictor, at 47%.

Table 5.14. Mean and median values of the ratio of the accuracy measures of EBP_{ρ} and the considered predictors and estimators of the domain totals ($\rho = -0.65$)

	$\frac{MSE(EBP_{\rho})}{MSE(.)}$		$\frac{RMSE(EBP_{\rho})}{RMSE(.)}$		$\frac{rRMSE(EBP_{\rho})}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_0	1.01	1.02	1.00	1.01	1.00	1.01
BP_{ρ}	1.62	1.64	1.27	1.28	1.27	1.28
$PLUGIN_{\rho}$	1.26	1.24	1.12	1.11	1.12	1.11
$PLUGIN_0$	1.28	1.25	1.13	1.12	1.13	1.12
HT	0.39	0.10	0.46	0.32	0.46	0.32
SYN	0.18	0.18	0.42	0.43	0.42	0.43
$CALIB$	0.32	0.11	0.43	0.34	0.43	0.34

Source: Own elaboration.

Table 5.15 contains the average values and median of the ratio of the accuracy measures of $PLUGIN_{\rho}$ and the other considered predictors and estimators of the domain totals. Comparing $PLUGIN_{\rho}$ with the other predictors considered in the simulation study, the largest average gain in accuracy was obtained relative to the predictors of the EBP class of approximately 10%.

It should be added that the gain values relative to the predictor EBP_0 are 5 percentage points higher than the results obtained in section 5.2.2. However, the highest values of average gain were observed when comparing the analysed estimators. For the synthetic estimator, this gain is approximately 40%. In the case of precision, the maximum average gain when compared with the estimators was close to the gain in precision, with a gain of approximately 25% when compared with EBP.

Table 5.15. Mean and median values of the ratio of the accuracy measures of $PLUGIN_\rho$ and the considered predictors and estimators of the domain totals ($\rho = -0.65$)

	$\frac{MSE(PLUGIN_\rho)}{MSE(.)}$		$\frac{RMSE(PLUGIN_\rho)}{RMSE(.)}$		$\frac{rRMSE(PLUGIN_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_ρ	0.79	0.81	0.89	0.90	0.89	0.90
EBP_0	0.80	0.79	0.89	0.89	0.89	0.89
BP_ρ	1.28	1.27	1.13	1.12	1.13	1.12
$PLUGIN_0$	1.01	1.01	1.00	1.00	1.00	1.00
HT	0.30	0.08	0.41	0.29	0.41	0.29
SYN	0.14	0.14	0.37	0.38	0.37	0.38
$CALIB$	0.25	0.09	0.38	0.30	0.38	0.30

Source: Own elaboration.

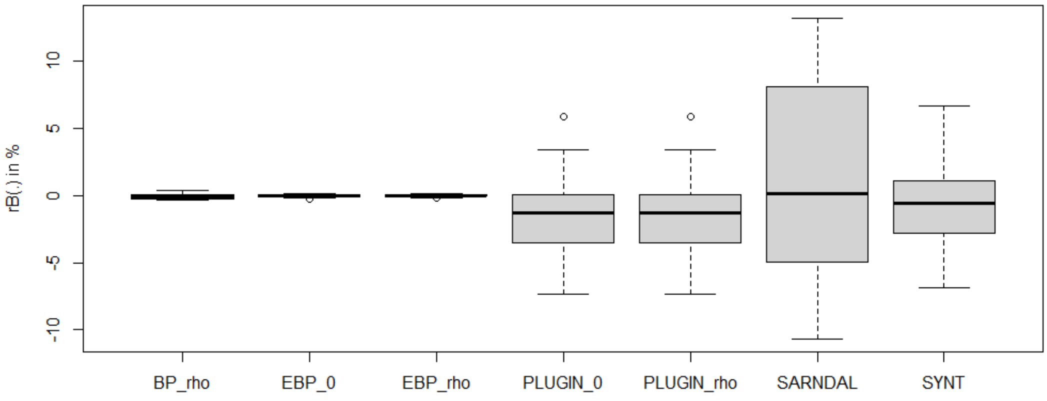


Figure 5.22. Values of $rB(.)$ of predictors and estimators of the median in domains ($\rho = -0.65$)

Source: Own elaboration.

Figures 5.22 and 5.23 show as box plots the relative bias values of the predictors and estimators of the second parameter analysed – the median in domain. For the proposed predictors BP_ρ and EBP_ρ, as in the other simulation analyses, including subsection 5.2.2, results as to the modulus were obtained close to 0. For the last of the proposals, i.e. the plug-in class predictor that takes into account the correlation between random effects and the synthetic estimator of the median in domain proposed by Stachurski (2018), the results are similar and their absolute value does not exceed 7%. In contrast, the highest $rB(.)$ modulus values were obtained for the estimator proposed by Särndal et al. (1992). They amount to as much as several percent.

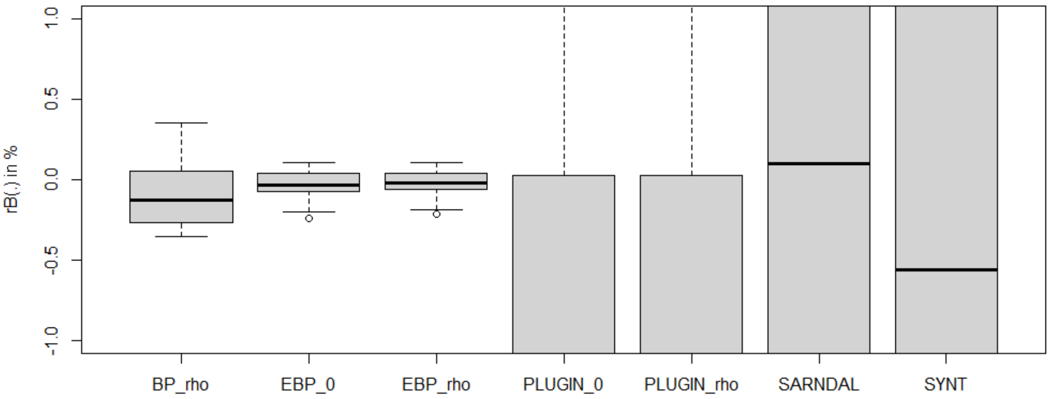


Figure 5.23. Selected $rB(.)$ values of predictors and estimators of the median in domains ($\rho = -0.65$)

Source: Own elaboration.

Figure 5.24 presents the $rRMSE(.)$ values of the predictors and estimators of the median in domains. For the BP- and EBP-class predictors, the relative $RMSE(.)$ values did not exceed 5%. It should be added that the median $rRMSE(.)$ for the proposed predictor BP_ρ was below 3%. For plug-in predictors, however, the maximum value of this measure was 7.90%. These results are therefore significantly similar to those obtained in the first variant of the study. The highest $rRMSE(.)$ values were obtained for the estimator presented by Särndal et al. (1992), for which the median value is about 9.5% and the maximum value over 14%.

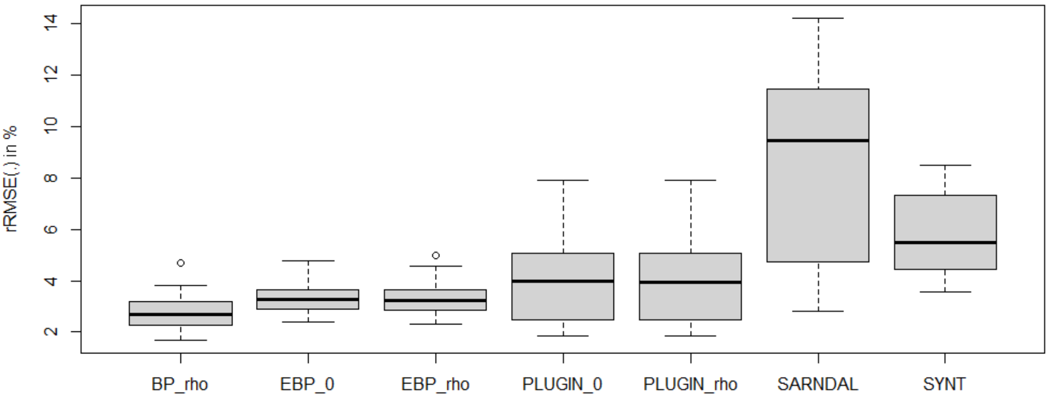


Figure 5.24. Values of $rRMSE(.)$ of predictors and estimators of the domains median ($\rho = -0.65$)

Source: Own elaboration.

For the relative $D(.)$ values of the predictors and the median estimators in the domain, the value of this measure did not exceed the level of 5% for all predictors, as in subsection 5.2.2. For the estimators, the maximum value of $rD(.)$ was even over 9%. For two of the proposed predictors, under the assumption of a model with correlated random effects (BP_ρ and $PLUGIN_\rho$), the median $rD(.)$ was below 3%.

Table 5.16 presents the mean and median values of the ratio of the accuracy measures of BP_ρ and the domains' median predictors and estimators considered in the simulation study. The recorded average gain in accuracy for BP_ρ relative to the two EBP-class predictors analysed is $17-18\%$. For the predictors $PLUGIN_\rho$ and $PLUGIN_0$, the median accuracy gain was less than 20%. The largest median accuracy gain was observed for the estimator of Särndal et al. (1992), at 61%. These results are similar to those obtained in the first variant of the study. In the case of precision, gains of a dozen and tens of percent were observed for the EBP and the estimators.

Table 5.16. Mean and median values of the ratio of the accuracy measures of BP_ρ and the considered predictors and estimators of the median in domains ($\rho = -0.65$)

	$\frac{MSE(BP_\rho)}{MSE(.)}$		$\frac{RMSE(BP_\rho)}{RMSE(.)}$		$\frac{rRMSE(BP_\rho)}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_ρ	0.69	0.71	0.83	0.84	0.83	0.84
EBP_0	0.68	0.70	0.82	0.83	0.82	0.83
$PLUGIN_\rho$	0.61	0.66	0.76	0.81	0.76	0.81
$PLUGIN_0$	0.61	0.65	0.76	0.81	0.76	0.81
$SARN$	0.18	0.11	0.39	0.34	0.39	0.34
SYN	0.26	0.25	0.50	0.50	0.50	0.50

Source: Own elaboration.

Table 5.17 presents the mean and median values of the ratio of the accuracy measures of EBP_ρ and the other predictors considered, as well as the median estimators in the domains. As in subsection 5.2.2 comparing EBP_ρ with the EBP-class predictor, which does not take into account correlations between random effects, an average gain in accuracy can be observed. By comparing the results for the proposed EBP predictor and the plug-in predictors, an approx. 7% average gain in accuracy can be noted. However, the highest values of median and average gain can be observed for the analysed estimators – of the order of several tens of percent. Gains in precision were also recorded.

Table 5.17. Mean and median values of the ratio of the accuracy measures of EBP_{ρ} and the considered predictors and estimators of the median in domains ($\rho = -0.65$)

	$\frac{MSE(EBP_{\rho})}{MSE(.)}$		$\frac{RMSE(EBP_{\rho})}{RMSE(.)}$		$\frac{rRMSE(EBP_{\rho})}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_0	0.99	0.99	0.99	0.99	0.99	0.99
BP_{ρ}	1.48	1.41	1.21	1.19	1.21	1.19
$PLUGIN_{\rho}$	0.94	0.99	0.93	0.99	0.93	0.99
$PLUGIN_0$	0.94	0.97	0.93	0.98	0.93	0.98
$SARN$	0.27	0.16	0.47	0.40	0.47	0.40
SYN	0.38	0.38	0.60	0.62	0.60	0.62

Source: Own elaboration.

Table 5.18 shows the mean and median values of the ratio of the accuracy measures of $PLUGIN_{\rho}$ and the considered predictors and estimators of the median in domains. The highest accuracy gain values can be seen when comparing the predictor $PLUGIN_{\rho}$ with the considered estimators. For both the estimator of Särndal et al. (1992) and the synthetic median-in-domain estimator proposed by Stachurski (2018), the average gain is approximately 40%, which is only three percentage points lower than in subsection 5.2.2. Similar results were also obtained for precision. Furthermore, juxtaposing the proposed plug-in predictor with the EBP-class predictors, an average gain of several percent was obtained.

Table 5.18. Mean and median values of the ratio of the accuracy measures of $PLUGIN_{\rho}$ and the considered predictors and estimators of the median in domains ($\rho = -0.65$)

	$\frac{MSE(PLUGIN_{\rho})}{MSE(.)}$		$\frac{RMSE(PLUGIN_{\rho})}{RMSE(.)}$		$\frac{rRMSE(PLUGIN_{\rho})}{rRMSE(.)}$	
	mean	median	mean	median	mean	median
EBP_{ρ}	1.69	1.04	1.21	1.02	1.21	1.02
EBP_0	1.69	1.05	1.21	1.02	1.21	1.02
BP_{ρ}	2.37	1.53	1.45	1.24	1.45	1.24
$PLUGIN_0$	1.00	1.00	1.00	1.00	1.00	1.00
$SARN$	0.48	0.24	0.58	0.48	0.58	0.48
SYN	0.50	0.47	0.69	0.69	0.69	0.69

Source: Own elaboration.

In summary, as in the other variants of the study, for the proposed EBP- and BP-class predictors, the relative values of the simulation bias were close to 0. The simulation-derived $rRMSE(.)$ values of these predictors were no higher than 4%, and for the plug-in predictor proposal, they did not exceed 8%. When comparing the properties of the predictor proposals with the considered estimators, analogous to the other analyses, an average gain in accuracy and precision of up to several tens of percent was noted. The maximum was over 90% in the case of accuracy and about 70% in terms of precision for the selected domains. When set against the selected predictors, the gain was lower.

5.5. Summary

This chapter was focused on the simulation studies carried out for the purpose of this book, conducted according to the model-based approach. Section 5.1 presented the considered dataset. The sampling assumptions and the considered division of the population into domains are also presented.

The following three subsections showed the assumptions and results of the analyses carried out. Subsection 5.2 was focused on the first variant of the simulation study. This section presented the algorithm according to which the analyses were carried out in order to compare the properties of selected statistics for assessing characteristics in domains. The predictors taken into account in the study are also discussed, among which are the three authors' proposed predictors, taking into account the correlation between the random effects of the BP, EBP and plug-in classes, respectively. In addition, selected predictors that assume a linear mixed model with uncorrelated random effects are included in the analyses. The list was also supplemented with some estimators of the considered characteristics belonging to the direct, indirect and calibrated groups. It should be added that the paper addresses the problem of estimating two characteristics: total values and medians in the domain. Among the quantities that allow an analysis of the properties of the above methods of estimating domain characteristics were measures of prediction precision, prediction accuracy and predictor bias.

The first variant of the simulation study assumed in the data generation process the use of the actual values of the auxiliary variable and the population-based estimated parameters of the random effects correlation model, including $\rho = -0.83$. The other two studies modified these assumptions. The study presented in subsection 5.3 considered the case of a stronger negative correlation ($\rho = -0.95$) and in 5.4 a weaker negative correlation ($\rho = -0.65$). Decisions on the modifications considered were based on the preliminary studies conducted. For each of the studies carried out, the simulation relative bias for the proposed BP- and EBP-class predictors were close to 0. In the case of the third proposal – the plug-in predictor – they did not exceed 10%, even assuming a weak correlation. The simulation-derived $rRMSE(.)$ values of the predictors taking into account random effects correlation did not exceed the 4.5% level for the estimated domain characteristics, except for the median prediction using the plug-in predictor, where the values were no higher than 8%. Furthermore, in the case of relative $D(.)$ values, the simulation maximum value of this measure also did not exceed a few percent.

Comparing the accuracy and precision of the EBP predictor, which both does and does not take into account correlations between random effects, the maximum average gain is 1% and 2%, respectively, when the characteristics of interest are the total value and the median. However, for individual domains, the maximum gain was as high as 5% and 7% for the above two characteristics. Comparing the properties of the proposed BP predictor and the EBP predictors, one can observe a maximum average gain in precision as well as accuracy of about 21% when the total value was the analysed characteristic and 18% when the median in terms of domains. The maximum gain in terms of domains, nevertheless, was even approximately 30% on accuracy and 34% on precision for the total value. When median prediction in domains was considered, the maximum gain in both cases was 30%. Comparing the accuracy and precision measures for the plug-in predictor proposal and the predictor of this class, assuming a model with uncorrelated random effects, the maximum average gain was 1% for both the total and median value. By contrast, when considering the problem in terms of domains, this maximum gain was twice as high for the total value and four times as high for the median across domains.

The results obtained may suggest a significant impact of the accuracy of parameter estimation, including the ρ parameter, on the results obtained. This is indicated, among other things, by comparing the gains in accuracy and precision of the proposed BP predictor and EBP against its counterpart, which does not take into account correlations between random effects. A possible solution to this problem in future studies is to include other iterative algorithms in the REML method used in the model estimation process or other methods for determining model parameter estimates. For the considered overpopulation model, the correlation between domain-specific random effects was taken into account – considering a larger number of domains into which the population is divided in future simulation studies could also significantly improve the accuracy of model parameter estimates (including the correlation coefficient). In the case of the plug-in predictor proposal, on the other hand, some modifications to reduce the bias on this predictor could also be considered in future analyses. However, a significant advantage of this statistic over EBP-class predictors is that it is less time-consuming to calculate.

It should be noted that when comparing the properties of the predictor proposals with the estimators considered in the study, one can see an average gain in accuracy and precision of up to several tens of percent regardless of the value of the ρ parameter. In contrast, for the selected domains, the maximum gain in precision reached up to 94% when the prediction of the total value was considered, and 84% when the median in domains was considered. The maximum gain in precision that was observed was 70% when the characteristic of interest was the total value, and 57% when the median in domains.

The chapter used the author's sections of code written in R to calculate the BP and EBP predictors of the total value and the median in domain under the assumption of a linear mixed model with correlated random effects. In addition, the author prepared code for conducting simulation studies.

Conclusions

The book presented the model-based approach in small area estimation and its applications in economic research, including the author's prediction methods. The monograph consisted of a theoretical and cognitive part – the first four chapters – and an empirical part (results of simulation studies – chapter 5).

Chapter one discussed the main approaches in small area estimation and their applications in research of an economic nature. Particular attention was given to the model-based approach, including the process of constructing overpopulation models and their classification. As part of this, generalisations of selected predictors to cross-sectional-temporal analyses were proposed. In addition, the author's proposals for some special cases of linear mixed models taking into account the correlation between random effects vectors and examples of applications in small area estimation are presented. For the above class of models, the use of permutation tests and permutation versions of classical tests in the verification of parameter significance was proposed. The problem was also supplemented with an author's proposal of a test allowing verification of the presence of correlations between vectors of random effects, based on the parametric bootstrap method.

Chapter two presented the issue of single and longitudinal surveys, their classification and applications. However, the issue of repeated surveys over time, including panel surveys, is discussed in more detail. Presented were, i.a., schemes of conducting them, advantages and disadvantages, as well as examples of research conducted in multiple periods in Poland and worldwide.

Chapter three was dedicated to the BLUPs and EBLUPs classes. The topics were presented in the light of the classification of linear mixed models. The author's proposal for the use of an empirical best linear unbiased predictor under the assumption of a linear mixed model with correlated random effects vectors is also presented. The chapter also addressed the problem of estimating the mean squared error of EBLUPs class, including a proposed modification of known methods for the above predictor proposal. In addition, selected modifications of the presented classes of predictors and their applications in economic research were discussed.

Chapter four presented the theoretical aspects of two classes of predictors – EBP and plug-in. The author's proposal to use these predictors in economic research assuming linear mixed models with correlated random effects vectors was also presented. The problem of estimating the mean squared errors of the EBP and plug-in class predictors is also addressed, including suggestions for modifying known methods for assessment the MSE of the proposed predictors. The chapter also discussed selected applications of the statistics considered.

The first four chapters provided answers to the first two research questions posed in this book. Chapter five, however, answered the third and fourth questions. It also made it possible to achieve the practical objectives of the monograph.

Chapter five contains a description of the dataset analysed, the assumptions, and the results of the simulation studies, conducted according to the model-based approach. The analyses were aimed at a simulation comparison of the properties of the author's proposed predictors of characteristics in domains, discussed in chapters three and four, with corresponding predictors that do not take into account correlations between the random effects vectors and the selected estimators. The characteristics considered in the prediction process were the total value and the median in the domain. Three variants of the simulation study were considered in this chapter, which took into account the different strength of the correlation between the random effects vectors at the population data generation stage. It should be added that the first variant assumed the original parameter values obtained from the real dataset considered. Comparisons of the properties of the above statistics for assessing characteristics in domains were made using measures of predictor accuracy and precision and predictor bias. The study was conducted using the R language (R Core Team, 2022) and custom-written functions. The results obtained suggest good properties of the author's considered predictor proposals, as indicated by the low simulated relative values of prediction standard error and root mean square error of prediction. The juxtaposition of the results obtained for the considered selected statistics for assessment characteristics in domains (predictors and estimators) suggests a gain in prediction accuracy and precision resulting from the application of the presented predictor proposals.

It should be emphasised that the results obtained in this study may be useful in practice for institutions conducting research as well as for the users of the data obtained as a result, e.g. state administration agencies at both the central and local levels. The analyses carried out in this study may form the basis for further research. These considerations may include the problem of the influence of the accuracy of the estimation of the parameters of the proposed models on the prediction process and the use of methods and algorithms other than those presented for estimating the parameters of the overpopulation model. The issue of the influence of the number of domains on the properties of the predictor proposals may also be included. It is also possible to expand the considered variants of simulation studies to include other models and distributions of effects and random components in the process of generating population data. The topics of this paper dealt with the model-based approach in small area estimation, but in further research, it would be possible to broaden the considerations to include an analysis of the properties of the proposed predictors conducted according to the design-based approach.

Bibliography

Act of 13 November 2003 on revenues of local government units (Journal of Laws of the Republic of Poland, 2021, 1672, consolidated text).

Act of 9 June 2011 – Geological and Mining Law (Journal of Laws of the Republic of Poland, 2021, 1420, consolidated text).

Aigner D.J., Hsiao C., Kapteyn A., Wansbeek T. (1984), *Latent variable models in econometrics* [in:] *Handbook of econometrics,* Z. Griliches, M.D. Intriligator (Eds.), 2, Amsterdam, 1321–1393.

Aitchison J. (1986), *The statistical analysis of compositional data*, Chapman and Hall, London.

Akaike H. (1973), *Maximum likelihood identification of Gaussian autoregressive moving average models*, "Biometrika", 60, 2, 255–265, https://doi.org/10.2307/2334537

Anderson T.W. (1959), *On asymptotic distributions of estimates of parameters of stochastic difference equations*, "The Annals of Mathematical Statistics", 30, 3, 676–687, https://www.jstor.org/stable/2237408

Andres H.J., Golsch K., Schmidt A.W. (2013), *Applied panel data analysis for economic and social surveys*, Springer Science & Business Media, Heidelberg-Berlin.

Antal E., Tille Y. (2014), *A new resampling method for sampling designs without replacement: The doubled half bootstrap*, "Computational Statistics", 29, 1345–1363, https://doi.org/10.1007/s00180-014-0495-0

Arellano M., Bover O., Labeaga J. (1999), *Autoregressive models with sample selectivity for panel data* [in:] *Analysis of panels and limited dependent variable models*, C. Hsiao, K. Lahiri, L.F. Lee, M.H. Pesaran (Eds.), Cambridge University Press, Cambridge, 23–48, https://www.researchgate.net/publication/5061545

Ashenfelter O., Solon G. (1982), *Longitudinal labor market data* [in:] *What's happening to American labor force and productivity measurements?* Proceedings of a June 17, 1982, Conference, sponsored by The National Council on Employment Policy. W.E. Upjohn Institute for Employment Research, Kalamazoo, MI, 109–126, https://doi.org/10.17848/9780880994149.ch6

Baldermann C., Salvati N., Schmid T. (2018), *Robust small area estimation under spatial non-stationarity*, “International Statistical Review”, 86, 1, 136–159, https://doi.org/10.1111/insr.12245

Balicki A. (1986), *Statystyczne metody badania płynności kadr: mierzenie i modelowanie*, Rozprawy i Monografie, 69, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.

Balicki A. (1989), *Wybrane problemy badań reprezentacyjnych w światowym piśmiennictwie statystycznym ostatnich lat* [in:] *Problemy badań statystycznych metodą˛ reprezentacyjną Konferencja naukowa Zwartowo 26-28 1988 r.*, R. Zasępa (Ed.), GUS, Warszawa, 30–45.

Balicki A. (1997), *Zastosowanie metod analizy kohortowej do badania czasu pozostawania bez pracy*, “Wiadomości Statystyczne”, 7, 53–62.

Bartosiewicz S. (1973), *Prosta metoda wyboru zmiennych objaśniających w modelu ekonometrycznym*, “Prace Naukowe WSE we Wrocławiu”, 43, 93–101.

Bartosińska D. (2008), *Analiza porównawcza gospodarstw rolnych*, “Wiadomości Statystyczne”, 12, 11–26.

Battese G.E., Harter R.M., Fuller W.A. (1988), *An error-components model for prediction of county crop areas using survey and satellite data*, “Journal of the American Statistical Association”, 83, 401, 28–36, https://doi.org/10.2307/2288915

Beaule A., Campbell F., Dascola M., Insolera N., Johnson D., Juska P., McGonagle K., Warra J. (2017), *PSID Main Interview User Manual: Release 2017*, Institute for Social Research, University of Michigan, Ann Arbor, MI, https://psidonline.isr.umich.edu/data/Documentation/UserGuide2017.pdf

Beenstock M., Felsenstein D. (2008), *Testing spatial stationarity and spatial cointegration*, Mimeo.

Benedetti R., Pratesi M., Salvati N. (2012), *Local stationarity in small-area estimation models*, “Statistical Methods and Applications”, 22, 1, 81-95, https://doi.org/10.1007/s10260-012-0208-1

Berg E., Chandra H. (2014), *Small area prediction for a unit-level lognormal model*, “Computational Statistics and Data Analysis”, 78, 159–175, https://doi.org/10.1016/j.csda.2014.03.007

Binder D.A., Hidiroglou M.A. (1988), *Sampling in time* [in:] *Sampling*, P.R. Krishnaiah, C.R. Rao (Eds.), Handbook of Statistics, 6, Elsevier, Amsterdam, 187–211.

Biecek P. (2012), *Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*, Wydawnictwo Naukowe PWN, Warszawa.

Biorn E. (1992), *Econometrics of panel data with measurement errors* [in:] *The econometrics of panel data: Theory and applications*, L. Matyas, P. Sevestre (Eds.), 1st ed., Kluwer Academic, Dordrecht, 152–195, https://doi.org/10.1007/978-94-009-0137-7_10

Bishop Y.M.M., Fienberg S.E., Holland P.W. (1975), *Discrete multivariate analysis: Theory and practice*, MIT Press, Cambridge, MA.

Booth A., Johnson D.R. (1985), *Tracking respondents in a telephone interview panel selected by random digit dialing*, "Sociological Methods & Research", 14, 1, 53–64, https://doi.org/10.1177/0049124185014001003

Borenstein M., Larry V., Hedges L.V., Higgins J.P.T., Rothstein H.R. (2010), *A basic introduction to fixed-effect and random-effects models for meta-analysis*, "Research Synthesis Methods", 1, 97–111, https://doi.org/10.1002/jrsm.12

Boubeta M., Lombardia M.J., Morales D. (2016), *Empirical best prediction under area-level Poisson mixed model*s, "Test", 25, 548–569, https://doi.org/10.1007/s11749-015-0469-8

Boubeta M., Lombardia M.J., Morales D. (2017), *Poisson mixed models for studying the poverty in small areas*, "Computational Statistical Data Analysis", 107, 32–47, https://doi.org/10.1016/j.csda.2016.10.014

Box G.E., Tiao G.C. (1968), *Bayesian estimation of means for the random effect model*, "Journal of the American Statistical Association", 63, 321, 174–181.

Bozdogan H. (1987), *Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions*, "Psychometrika", 52, 345–370.

Bracha Cz. (1994), *Metodologiczne aspekty badania małych obszarów*, „Z Prac Zakładu Badań Statystyczno–Ekonomicznych", z. 43, ZBSE GUS i PAN, Warszawa.

Bracha Cz. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, PWN, Warszawa.

van den Brakel J.A., Buelens B., Boonstra H.-J. (2016), *Small area estimation to quantify discontinuities in repeated sample surveys*, "Journal of the Royal Statistical Society, Series A (Statistics in Society)", 179, 1, 229–250, https://doi.org/10.1111/rssa.12110

van den Brakel J.A., Krieg S. (2016), *Small area estimation with state space common factor models for rotating panels*, "Journal of the Royal Statistical Society, Series A (Statistics in Society)", 179, 3, 763–791, https://www.jstor.org/stable/43965818

Bramati M.C., Croux C. (2007), *Robust estimators for the fixed effects panel data model*, "The Econometrics Journal", 10, 3, 521–540, https://www.jstor.org/stable/23126789

Brewer K.R.W. (1995), *Combining design-based and model-based inference* [in:] *Business survey methods*, B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, P.S. Kott (Eds.), Wiley, New York, 589–606, https://doi.org/10.1002/9781118150504.ch30

Brewer K.R.W. (1999), *Design-based or prediction-based inference? Stratified random vs stratified balanced sampling*, "International Statistical Review", 67, 1, 35–47, https://doi.org/10.2307/1403564

Brewer K.R.W., Hanif M., Tam S.M. (1988), *How nearly can model-based prediction and design-based estimation be reconciled?* "Journal of the American Statistical Association", 83, 401, 128–132, https://doi.org/10.2307/2288930

Britsish Social Attitiudes Survey (2016), *User guide*, NatCen Social Research that Works for Society, London.

Brunsdon C., Fotheringham A.S., Charlton M. (1996), *Geographically weighted regression: A method for exploring spatial nonstationarity*, "Geographical Analysis", 28, 4, 281–298, https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Butar F.B., Lahiri P. (2003), *On measures of uncertainty of empirical Bayes small-area estimators*, "Journal of Statistical Planning and Inference", 112, 63–76, https://www.math.umd.edu/~plahiri/pdfs/ButarLahiriJSPI2003.pdf

Call V.R.A., Otto L.B., Spenner K.I. (1982), *Tracking respondents: A multi-method approach*, Lexington Books, Lexington, Mass.

Cassel C.M., Sarndal C.E., Wretman J.H. (1977), *Foundations of inference in survey sampling*, John Wiley & Sons, New York.

Chambers R., Chandra H., Salvati N., Tzavidis N. (2014), *Outlier robust small area estimation*, "Journal of the Royal Statistical Society", Series B, 76, 1, 47–69, https://www.jstor.org/stable/24772745

Chambers R., Chandra H., Tzavidis N. (2011), *On bias-robust mean squared error estimation linear small area estimators*, "Survey Methodology", 37, 153–170, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11604-eng.pdf

Chambers R., Salvati N., Tzavidis N. (2016), *Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK*, "Journal of the Royal Statistical Society", Series A, 179, 2, 453–479, https://www.jstor.org/stable/43965552

Chambers R.L., Chandra H. (2006), *Improved direct estimators for small areas*, Methodology Working Paper, M06/07, Southampton Statistical Science Research Institute, University of Southampton, U.K., https://eprints.soton.ac.uk/466510/1/1202326.pdf

Chandra H., Bathla H.V.L., Sud U.C. (2010), *Small area estimation under a mixture model*, "Statistics in Transition – New Series", 11, 3, 503–516, https://www.infona.pl/resource/bwmeta1.element.desklight-c1cff48a-cd6d-474b-9766-c9a7de71ce67

Chandra H., Chambers R. (2005), *Comparing EBLUP and C-EBLUP for small area estimation*, "Statistics in Transition", 7, 637–648.

Chandra H., Chambers R. (2006), *Small area estimation for skewed data*, "Southampton Statistical Sciences Research Institute, MethodologyWorking Papers", M06/05, University of Southampton, U.K.

Chandra H., Chambers R., Salvati N. (2012a), *Small area estimation of proportions in business surveys*, "Journal of Statistical Computation and Simulation", 82, 6, 783–795, https://ro.uow.edu.au/cgi/viewcontent.cgi?article=1034&context=cssmwp

Chandra H., Kumar S., Aditya K. (2018), *Small area estimation of proportions with different levels of auxiliary data*, "Biometrical Journal", 60, 2, 395–415, https://doi.org/10.1002/bimj.201600128

Chandra H., Salvati N., Chambers R., Tzavidis N. (2012b), *Small-area estimation under spatial non-stationarity*, "Computational Statistics and Data Analysis", 56, 2875–2888, https://doi.org/10.1016/j.csda.2012.02.006

Chatterjee S., Lahiri P., Li H. (2008), *Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models*, "Annals of Statistics", 36, 3, 1221–1245, https://www.jstor.org/stable/25464665

Chaudhuri A., Stenger H. (2005), *Survey sampling: Theory and methods*, CRC Press, Boca Raton, https://doi.org/10.1201/9781420028638

Chen J., Liu Y. (2019), *Small area quantile estimation*, "International Statistical Review", 87, S1, 219–238, https://doi.org/10.1111/insr.12293

Chen S., Lahiri P. (2002), *A weighted jackknife MSPE estimator in small-area estimation*, "Proceedings of the Section on Survey Research Methods", 473–477, http://www.asasrms.org/Proceedings/y2002/Files/JSM2002-001127.pdf

Chen S., Lahiri P. (2003), *A comparison of different MSPE estimators of EBLUP for the Fay–Herriot model*, "Proceedings of the Section on Survey Research Methods", 905–911, http://www.asasrms.org/Proceedings/y2003/Files/JSM2003-000585.pdf

Choi J., Fuentes M., Reich B.J. (2009), *Spatial-temporal association between fine particulate matter and daily mortality*, "Computational Statistics Data Analysis", 53, 8, 2989–3000, https://doi.org/10.1016/j.csda.2008.05.018

Chwila A., Żądło T. (2019), *On properties of empirical best predictors*, "Communications in Statistics – Simulation and Computation", 51, 1, 1–34, https://doi.org/10.1080/03610918.2019.1649422

Cochran W.G. (1939), *The use of the analysis of variance in enumeration by sampling*, "Journal of the American Statistical Association", 34, 207, 492–510, https://doi.org/10.1080/01621459.1939.10503549

Conley T.G. (1999), *GMM estimation with cross sectional dependence*, "Journal of Econometrics", 92, 1, 1–45, https://doi.org/10.1016/S0304-4076(98)00084-0

Courgeau D., Lelievre E. (1988), *Analyse demographique des biographies*, INED, Paris.

Cox D.R., Hinkley D.V. (1974), *Theoretical statistics*, Chapman & Hall, London.

Cressie N. (1993), *Statistics for spatial data*, John Wiley & Sons, Hoboken, NJ, https://doi.org/10.1002/9781119115151

Czapiński J., Panek T., Eds. (2000), *Diagnoza społeczna. Warunki i jakość życia Polaków oraz ich doświadczenia z reformami systemowymi po 10 latach transformacji*, Rada Monitoringu Społecznego i Wyższa Szkoła Pedagogiczna TWP, Warszawa, http://www.diagnoza.com/pliki/raporty/Diagnoza_raport_2000.pdf

Czapiński J., Panek T., Eds. (2015), *Diagnoza społeczna 2015. Warunki i jakość życia Polaków*, Rada Monitoringu Społecznego, Warszawa, http://www.diagnoza.com/pliki/raporty/Diagnoza_raport_2015.pdf

Dacey M. (1968), *A review of measures of contiguity for two and k-color maps* [in:] *Spatial analysis: A reader in statistical geography*, B. Berry, D. Marble (Eds.), Prentice-Hall, Englewood Cliffs, N.J., 479–495.

Datta G.S., Lahiri P. (2000), *A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems*, "Statistica Sinica", 10, 613–627, https://www.jstor.org/stable/24306735

Datta G.S., Rao J.N.K., Smith D.D. (2005), *On measuring the variability of small area estimators under a basic area level model*, "Biometrika", 92, 1, 183–196, https://www.jstor.org/stable/20441175

Dehnel G. (2003), *Statystyka małych obszarów jako narzędzie oceny rozwoju ekonomicznego regionów*, Wydawnictwo Akademii Ekonomicznej, Poznań.

Dehnel G. (2010), *Estymacja odporna a efektywność szacunku na podstawie badania mikroprzedsiębiorstw*, "Zeszyty Naukowe – Uniwersytet Ekonomiczny w Poznaniu", 149, 162–178, https://bazekon.uek.krakow.pl/zeszyty/171205913

Dehnel G. (2018), *Dobór modelu a obciążenie szacunku na przykładzie estymatora GREG w badaniu małych przedsiębiorstw*, "Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie", 11, 971, 5–25, https://doi.org/10.15678/ZNUEK.2017.0971.1101

Demidenko E. (2004), *Mixed models: Theory and application*, John Wiley & Sons, Hoboken, https://doi.org/10.1002/0471728438

Dempster A.P., Rubin D.B., Tsutakawa R.K. (1981), *Estimation in covariance components models*, "Journal of the American Statistical Association", 76, 374, 341–353, https://doi.org/10.2307/2287835

Deng Y., Hillygus D.S., Reiter J.P., Si Y., Zheng S. (2013), *Handling attrition in longitudinal studies: The case for refreshment samples*, “Statistical Science”, 28, 2, 238–256, https://doi.org/10.1214/13-STS414

Deville J.C., Sarndal C.-E. (1992), *Calibration estimators in survey sampling*, “Journal of the American Statistical Association”, 87, 418, 376–382, https://doi.org/10.2307/2290268

Diallo M.S., Rao J.N.K. (2018), *Small area estimation of complex parameters under unit-level models with skew-normal errors*, “Scandinavian Journal of Statistics”, 45, 4, 1092-1116, https://doi.org/10.1111/sjos.12336

Dickey D.A., Fuller W.A. (1979), *Distribution of the estimators for autoregressive time series with a unit root*, “Journal of the American Statistical Association”, 74, 366a, 427–431, https://doi.org/10.2307/2286348

Dickey D.A., Fuller W.A. (1981), *Likelihood ratio statistics for autoregressive time series with a unit root*, “Econometrica: Journal of the Econometric Society”, 49, 1057–1072, https://doi.org/10.2307/1912517

Dol W. (1991), *Small area estimation: A synthesis between sampling theory and econometrics*, Wolters-Noordhoff, Groningen.

Domański C., Pruska K. (1996), *Reprezentatywność próby w statystyce małych obszarów*, “Wiadomości Statystyczne”, 41, 5, 11-16.

Domański C., Pruska K. (1997), *Prognozowanie w przedsiębiorstwie z wykorzystaniem statystyki małych obszarów* [in:] *Prognozowanie w zarządzaniu firmą*, M. Cieślak (Ed.), Materiały konferencyjne, Wydawnictwo Akademii Ekonomicznej, Wrocław, 49–56.

Domański C., Pruska K. (2001), *Metody statystyki małych obszarów*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Dumont C., Chenel M., Mentre F. (2014), *Influence of covariance between random effects in design for nonlinear mixed-effect models with an illustration in pediatric pharmacokinetics*, “Journal of Biopharmaceutical Statistics”, 24, 3, 471–492, https://doi.org/10.1080/10543406.2014.888443

Duncan G.J., Juster F.T., Morgan J.N. (1986), *The role of panel studies in a world of scarce research resources* [in:] *Survey research designs: Towards a better understanding of their costs and benefits*, R.W. Pearson, R.F. Boruch (Eds.), Springer, New York, 94–129, https://link.springer.com/chapter/10.1007/978-1-4684-6336-1_5

Duncan G.J., Kalton G. (1987), *Issues of design and analysis of surveys across time*, “International Statistical Review”, 55, 1, 97–117, https://doi.org/10.2307/1403273

Eideh A.A.H., Nathan G. (2006), *Fitting time series models for longitudinal survey data under informative sampling*, “Journal of Statistical Planning and Inference”, 136, 3052–3069, https://bibliotekanauki.pl/articles/465750.pdf

Elbers Ch., van der Weide R. (2014), *Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality*, “Policy Research” Working Paper, World Bank Group, 6962, 1–31, https://documents1.worldbank.org/curated/en/712781468338974024/pdf/WPS6962.pdf

Esteban M.D., Lombardia M.J., Lopez-Vizcaino E., Morales D., Perez A. (2020), *Small area estimation of proportions under area-level compositional mixed models*, “Test”, 29, 793–818, https://doi.org/10.1007/s11749-019-00688-w

Esteban M.D., Morales D., Perez A., Santamaria L. (2012), Small area estimation of poverty proportions under area-level time models, “Computational Statistics and Data Analysis”, 56, 2840–2855, https://doi.org/10.1016/j.csda.2011.10.015

Fabrizi E., Salvati N., Pratesi M., Tzavidis N. (2014), *Outlier robust model-assisted small area estimation*, “Biometrical Journal”, 56, 1, 157–175, https://doi.org/10.1002/bimj.201200095

Falorsi P.D., Falorsi S., Russo A. (1998), *Small area estimation at provincial level in the Italian Labour Force Survey*, “Journal of the Italian Statistical Society”, 7, 1, 93–109, https://doi.org/10.1007/BF03178923

Fay III R.E., Herriot R.A. (1979), *Estimates of income for small places: An application of James-Stein procedures to census data*, “Journal of the American Statistical Association”, 74, 366a, 269–277, https://doi.org/10.1080/01621459.1979.10482505

Fellner W.H. (1986), *Robust estimation of variance components*, “Technometrics”, 28, 51–60, https://doi.org/10.2307/1269603

Ferrante M.R., Pacei S. (2004), *Small area estimation for longitudinal surveys*, “Statistical Methods and Applications”, 13, 3, 327–340, https://doi.org/10.1007/s10260-004-0082-6

Fitzmaurice G.M., Laird N.M., Ware J.H. (2004), *Applied Longitudinal Analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons.

Fisher R.A. (1922), *On the mathematical foundations of theoretical statistics*, “Philosophical Transactions of the Royal Society London”, A, 222, 309–368, https://doi.org/10.1098/rsta.1922.0009

Fotheringham A.S., Brunsdon C., Charlton M. (2002), *Geographically weighted regression*, John Wiley and Sons, UK, https://www.researchgate.net/publication/27246972

Frątczak E., Ed. (2012), *Zaawansowane metody analiz statystycznych*, Oficyna Wydawnicza SGH, Warszawa.

Fuentes M., Song H.R., Ghosh S.K., Holland D.M., Davis J.M. (2006), *Spatial association between speciated fine particles and mortality*, “Biometrics”, 62, 3, 855–863, https://doi.org/10.1111/j.1541-0420.2006.00526.x

Fuller W.A. (1976), *Introduction to statistical time series*, John Wiley & Sons, New York--London-Sydney.

Fuller W.A. (1990), *Analysis of repeated surveys*, “Survey Methodology”, 16, 2, 167–180, https://www150.statcan.gc.ca/n1/pub/12-001-x/1990002/article/14537-eng.pdf

Fuller W.A. (1991), *Simple estimators for the mean of skewed populations*, “Statistica Sinica”, 1, 137–158, https://www.jstor.org/stable/24303997

Ghosh M., Rao J.N.K. (1994), *Small area estimation: An appraisal*, “Statistical Science”, 9, 1, 55–93, https://www.jstor.org/stable/2246284

Giesecke D. (1989), *Die Auflosung von Familien und das Wohlfahrtsniveau geschiedener Frauen. Frauenerwerbstatigkeit-Berichte aus der laufenden Forschung*, “SAMF Working Paper”, 97–111.

Giusti C., Marchetti S., Pratesi M., Salvati N. (2012), *Semi-parametric Fay-Herriot model using penalized splines*, “Journal of the Indian Society of Agricultural Statistics”, 66, 1–14, https://www.researchgate.net/publication/230806596

Gołata E. (1995), *Konstrukcja tablic aktywności zawodowej ludności z uwzględnieniem bezrobotnych* [in:] *Rozwój metodologii badań statystycznych w Polsce*, Biblioteka Wiadomości Statystycznych, GUS, Warszawa.

Gołata E. (1996), *Statystyka małych obszarów w analizie rynku pracy*, “Wiadomości Statystyczne”, 3, 45–59.

Gołata E. (2004), *Estymacja pośrednia bezrobocia na lokalnym rynku pracy*, Prace Habilitacyjne, Wydawnictwo Akademii Ekonomicznej, Poznań.

Gonzalez M.E. (1973), *Use and evaluation of synthetic estimates*, Proceedings of the Social Statistics Section, “American Statistical Association”, 33–36, http://www.asasrms.org/Proceedings/y1973/Use%20And%20Evaluation%20Of%20Synthetic%20Estimates.pdf

Gonzalez M.E., Hoza C. (1978), *Small-area estimation with application to unemployment and housing estimates*, “Journal of the American Statistical Association”, 73, 361, 7–15.

Gonzalez-Manteiga W., Lombardia M., Molina I., Morales D., Santamaria L. (2007), *Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model*, “Computational Statistical Data Analysis”, 51, 2720–2733, https://doi.org/10.1016/j.csda.2006.01.012

Gonzalez-Manteiga W., Lombardia M.J., Molina I., Morales D., Santamaria L. (2008), *Bootstrap mean squared error of a small-area EBLUP*, “Journal of Statistical Computation and Simulation”, 78, 443–462, https://doi.org/10.1080/00949650601141811

Goodman A.C., Thibodeau T.G. (1998), *Housing market segmentation*, "Journal of Housing Economics", 7, 2, 121–143, https://doi.org/10.1006/jhec.1998.0229

Graf M., Marin J.M., Molina I. (2018), *A generalized mixed model for skewed distributions applied to small area estimation*, "TEST", 27, 69, 1–33, https://doi.org/10.1007/s11749-018-0594-2

Graf M., Marin J.M., Molina I. (2019), *A generalized mixed model for skewed distributions applied to small area estimation*, "TEST", 28, 565–597, https://doi.org/10.1007/s11749-018-0594-2

Granger C.W.J. (1990), *Aggregation of time-series variables: A survey* [in:] *Disaggregation in econometric modeling*, T. Barker, M.H. Pesaran (Eds.), Routledge, London.

Guzik B. (2008), *Podstawy ekonometrii*, Wydawnictwo Akademii Ekonomicznej, Poznań.

Hajek J. (1981), *Sampling from a finite population*, Marcel Dekker, New York.

Hall D.B., Clutter M. (2004), *Multivariate multilevel nonlinear mixed effects models for timber yield predictions*, "Biometrics", 60, 1, 16–24, https://doi.org/10.1111/j.0006-341x.2004.00163.x

Hall P., Maiti T. (2006), *On parametric bootstrap methods for small area prediction*, "Journal of the Royal Statistical Society", Series B, 68, 2, 221–238, https://www.jstor.org/stable/3647567

Hannan E.J., Quinn B.G. (1979), *The determination of the order of an autoregression*, "Journal of the Royal Statistical Society: Series B (Methodological)", 41, 2, 190–195, https://www.jstor.org/stable/2985032

Hartley H.O., Rao J.N.K. (1967), *Maximum-likelihood estimation for the mixed analysis of variance model*, "Biometrika", 54, 1–2, 93–108, https://doi.org/10.2307/2333854

Harville D.A., Jeske D.R. (1992), *Mean squared error of estimation or prediction under a general linear model*, "Journal of the American Statistical Association", 87, 724–731, https://doi.org/10.2307/2290210

Hellwig Z. (1969), *Problem optymalnego wyboru predyktant*, "Przegląd Statystyczny", 3–4, 221–238.

Henderson C.R. (1950), *Estimation of genetic parameters* (Abstract), "Annals of Mathematical Statistics", 21, 309–310, https://www.scirp.org/reference/referencespapers?referenceid=1741014

Henry K., Lahiri P., Scali J. (2009), *Using sample data to reduce nonsampling error in state-level estimates produced from tax records* [in:] *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington D.C., 3571–3579, http://www.asasrms.org/Proceedings/y2009f.html

Hermalin B.E., Wallace N.E. (2001), *Firm performance and executive compensation in the savings and loan industry*, “Journal of Financial Economics”, 61, 1, 139–170, https://doi.org/10.2139/ssrn.68980

Hobza T., Morales D. (2016), *Empirical best prediction under unit-level logit mixed models*, “Journal of Official Statistics”, 32, 3, 661–692, https://doi.org/10.1515/jos-2016-0034

Horvitz D.G., Thompson D.J. (1952), *A generalization of sampling without replacement from a finite universe*, “Journal of the American statistical Association”, 47, 260, 663–685.

Hsiao C. (1999), *Analysis of panel data*, Cambridge University Press, Cambridge.

Hsiao C. (2007), *Panel data analysis — advantages and challenges*, “TEST”, 16, 1, 1–22, https://doi.org/10.1007/s11749-007-0046-x

Hsiao C., Appelbe T.W., Dineen C.R. (1993), *A general framework for panel data models with an application to Canadian customer-dialed long distance telephone service*, “Journal of Econometrics”, 59, 1–2, 63–86, https://doi.org/10.1016/0304-4076(93)90039-8

Hsiao C., Mountain D.C., Chan M.L., Tsui K.Y. (1989), *Modeling Ontario regional electricity system demand using a mixed fixed and random coefficients approach*, “Regional Science and Urban Economics”, 19, 4, 565–587, https://doi.org/10.1016/0166-0462(89)90020-3

Huber P.J. (1964), *Robust estimation of a location parameter*, “The Annals of Mathematical Statistics”, 73–101, https://www.jstor.org/stable/2238020

Huggins R.M. (1993), *On the robust analysis of variance components models for pedigree data*, “Australian Journal of Statistics”, 35, 1, 43–57, https://doi.org/10.1111/j.1467-842X.1993.tb01311.x

Ing Ch.-K. (2004), *Selecting optimal multistep predictors for autoregressive processes of unknown order*, “The Annals of Statistics”, 32, 2, 693–722, https://www.jstor.org/stable/3448482

Jackowska B. (2015), *Analiza kohortowa czasu istnienia mikroprzedsiębiorstw w Gdańsku*, “Journal of Management and Finance”, 13, 4, 127–145, https://repozytorium.bg.ug.edu.pl/info/article/UOG6e78541370e14a91a95e569f40941d25/

Jacqmin-Gadda H., Sibillot S., Proust C., Molina J.-M., Thiebaut R. (2006), *Robustness of the linear mixed models to misspecified error distribution*, “Computational Statistics and Data Analisis”, 51, 5141–5154, https://doi.org/10.1016/j.csda.2006.05.021

Jędrzejczak A. (2011), *Metody analizy rozkładów dochodów i ich koncentracji*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Jędrzejczak A., Kubacki J. (2016), *Estimation of mean income for small areas in Poland using Rao-Yu model*, "Acta Universitatis Lodziensis Folia Oeconomica", 3, 322, 37–53, https://doi.org/10.18778/0208-6018.322.05

Jędrzejczak A., Kubacki J. (2017), *Estimation of small area characteristics using multivariate Rao-Yu model*, "Statistics", 725–742, https://www.econstor.eu/bitstream/10419/207883/1/10.21307_stattrans-2017-009.pdf

Jhun M., Song S.H., Jung B.C. (2003), *BLUP in the nested panel regression model with serially correlated errors*, "Computational Statistics & Data Analysis", 44, 1–2, 77–88, https://doi.org/10.1016/S0167-9473(02)00348-1

Jiang J. (1996), *REML estimation: Asymptotic behavior and related topics*, "The Annals of Statistics", 24, 255–286.

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, https://doi.org/10.1016/S0378-3758(02)00293-8

Jiang J. (2007), *Linear and generalized linear mixed models and their applications*, Springer, New York.

Jiang J., Lahiri P. (2006), *Mixed model prediction and small area estimation*, "TEST", 15, 1–96, https://www.math.umd.edu/~plahiri/pdfs/JiangLahiriTest06.pdf

Jiang J., Lahiri P., Wan S.-M. (2002), *Unified jackknife theory for empirical best prediction with M-estimation*, "The Annals of Statistics", 30, 1782–1810, https://www.jstor.org/stable/1558740

Jiang J., Nguyen T. (2012), *Small area estimation via heteroscedastic nested-error regression*, "The Canadian Journal of Statistics", 40, 3, 588–603, https://www.jstor.org/stable/41724546

Jiang J., Nguyen T., Rao J.S. (2011), *Best predictive small area estimation*, "Journal of the American Statistical Association", 106, 494, 732–745, https://www.jstor.org/stable/41416406

Jiang J., Tang E.-T. (2011), *The best EBLUP in the Fay-Herriot model*, "Annals of the Institute of Statistical Mathematics", 63, 1123–1140, https://doi.org/10.1007/s10463-010-0281-x

Kackar R.N., Harville D.A. (1981), *Unbiasedness of two-stage estimation and prediction procedures for mixed linear models*, "Communications in Statistics, Series A", 10, 1249–1261, https://doi.org/10.1080/03610928108828108

Kackar R.N., 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, https://doi.org/10.1080/01621459.1984.10477102

Kalton G. (2009), *Design for surveys over time* [in:] *Design, method and applications*, D. Pfeffermann, C.R Rao (Eds.),Handbook of Statistics, 29A, 89–108, Elsevier, Amsterdam.

Kalton G., Citro C.F. (1993), *Panel surveys: Adding the fourth dimension*, "Survey Methodology", 19, 2, 205–215, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993002/article/14452-eng.pdf?st=QQZADbhz

Kalton G., Kordos J., Płatek R. (1993), *Small area statistics and survey designs*, Vol. I: *Invited papers*; Vol. II: *Contributed papers and panel discussion*, Central Statistical Office, Warsaw.

Karlberg F. (2000), *Survey estimation for highly skewed populations in the presence of zeros*, "Journal of Official Statististics", 16, 3, 229–241, https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/survey-estimation-for-highly-skewed-populations-in-the-presence-of-zeroes.pdf

Karpuk M. (2015), *Wpływ czynników przestrzennych na ruch turystyczny w województwie zachodniopomorskim (2006-2012)*, "Zeszyty Naukowe Wydziału Nauk Ekonomicznych Politechniki Koszalińskiej", 1, 19, 39–56, https://ezeszyty.wne.tu.koszalin.pl/index.php/zeszyty/article/view/13/14

Kędzior Z., Ed. (2005), *Badania rynku. Metody, zastosowania*, PWE, Warszawa.

Kiaer A.N. (1897), *The representative method of statistical survey*, Central Bureau of Statistics of Norway, Oslo.

Kiersztyn A., Życzyńska-Ciołek D., Słomczyński K.M., Eds. (2017), *Rozwarstwienie społeczne. Zasoby, szanse i bariery*, IFIS PAN, Warszawa.

Klimanek T. (2012), *Using indirect estimation with spatial autocorrelation in social surveys in Poland*, "Przegląd Statystyczny", 59 (special edition 1), 155–172, https://bibliotekanauki.pl/articles/422834.pdf

Kordos J. (1992), *Podejścia do statystyki małych obszarów w Polsce*, "Wiadomości Statystyczne", 10, 1–5.

Kordos J. (1997), *40 lat badań budżetów gospodarstw domowych w Polsce*, "Wiadomości Statystyczne", 42, 7, 27–42.

Kordos J. (1999), *Problemy estymacji danych dla małych obszarów*, "Wiadomości Statystyczne", 1, 85–101.

Kowalczyk B. (2001), *Badania powtarzalne w czasie*, Monografie i Opracowania, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa.

Kowalczyk B. (2013), *Zagadnienia estymacji złożonej w badaniach reprezentacyjnych opartych na próbach rotacyjnych*, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa.

Kriegler B., Berk R. (2010), *Small area estimation of the homeless in Los Angeles: An application of cost-sensitive stochastic gradient boosting*, "The Annals of Applied Statistics", 1234–1255, https://doi.org/10.1214/10-AOAS328

Krzciuk M. (2018), *On verification of a superpopulation model*, Paper presented at the 2nd Congress of Polish Statistics, Warsaw 10–12 July 2018, https://kongres.stat.gov.pl/images/prezentacja/12/p2-3._m_krzciuk_on_verification.pdf

Krzciuk M. (2019), *On the properties of EBP for a class of models with correlated random effects and longitudinal data*, Paper presented at the Italian Confrence on Survey Methodology, Florence 5–7 June 2019.

Krzciuk M. (2020), *On empirical best linear unbiased predictor under a Linear Mixed Model with correlated random effects*, "Econometrics", 24, 2, 17–29, https://doi.org/10.15611/eada.2020.2.02

Krzciuk M., Stachurski T., Żądło T. (2017), *On empirical best predictors of poverty measures based on Polish household budget survey*, "Studia i Materiały. Miscellanea Oeconomicae", 3, t. I. *Pomiar jakości życia w układach regionalnych i krajowych: dylematy i wyzwania*, 33–44.

Krzciuk M., Żądło T. (2013), *O testach istotności parametrów liniowych modeli mieszanych w badaniach wielookresowych w pakiecie R*, "Rola informatyki w naukach ekonomicznych i społecznych. Innowacje i implikacje interdyscyplinarne", 2, 197–213.

Krzciuk M., Żądło T. (2014a), *On some tests of variance components for linear mixed models*, "Studia Ekonomiczne", 189, 77–85, https://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-6f9694b4-788e-432f-8a34-49d1180a6dbe/c/8_M.K.Krzciuk_T.Zadlo_On_some_tests_of_variance..._01.pdf

Krzciuk M., Żądło T. (2014b), *On some tests of fixed effects for linear mixed models*, "Studia Ekonomiczne", 189, 49–57, https://sbc.org.pl/publication/131538

Kubacki J. (1997), *Ważniejsze metody estymacji w statystyce małych obszarów*, "Wiadomości Statystyczne", 42, 5, 13–22.

Kuhn E., Lavielle M. (2005), *Maximum likelihood estimation in nonlinear mixed effects models*, "Computational Statistics & Data Analysis", 49, 4, 1020–1038, https://doi.org/10.1016/j.csda.2004.07.002

Lawson A.B., Choi J., Cai B., Hossain M., Kirby R.S., Liu J. (2012), *Bayesian 2-stage space-time mixture modeling with spatial misalignment of the exposure in small area health data*, "Journal of Agricultural, Biological, and Environmental Statistics", 17, 3, 417–441, https://doi.org/10.1007/s13253-012-0100-3

Lazarsfeld P.F. (1940), *"Panel" studies*, "The Public Opinion Quarterly", 4, 1, 122–128, https://doi.org/10.1086/265373

Lazarsfeld P.F., Berelson B., Gaudet H. (1944), *The people's choice. How the voter makes up his mind in a presidential campaign,* Duell, Sloan and Pearce, New York.

Lazarsfeld P., Fiske M. (1938), *The "panel" as a new tool for measuring opinion*, "Public Opinion Quarterly", 2, 4, 596–612, https://doi.org/10.1086/265234

Lewbel A. (1994), *Aggregation and simple dynamics*, "American Economic Review", 84, 905–918, https://www.jstor.org/stable/2118037

Li Y., Lahiri P. (2007), *Robust model-based and model-assisted predictors of the finite population total*, "Journal of the American Statistical Association", 102, 478, 664–673, https://doi.org/10.1198/016214507000000158

Lilliefors H.W. (1967), *On the Kolmogorov-Smirnov test for normality with mean and variance unknown*, "Journal of the American Statistical Association", 66, pp. 399–402, https://doi.org/10.2307/2283970

Lindstrom M.J., Bates D.M. (1990), *Nonlinear mixed effects models for repeated measures data*, "Biometrics", 673–687, https://doi.org/10.2307/2532087

Littell R.C., Milliken G.A., Stroup W.W., Wolfinger R.D., Schabenberger O. (2006), *SAS for mixed models,* Second ed., SAS Institute, Cary, NC.

Lohr S.L., Prasad N.N. (2003), *Small area estimation with auxiliary survey data*, "Canadian Journal of Statistics", 31, 4, 383–396, https://doi.org/10.2307/3315852

Lohr S.L., Rao J.N.K. (2009), *Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models*, "Biometrika", 96, 2, 457–468, https://doi.org/10.1093/biomet/asp003

Longford N.T. (2004), *Missing data and small area estimation in the UK Labour Force Survey,* "Journal of the Royal Statistical Society, Series A (Statistics in Society)", 167, 2, 341–373, https://www.jstor.org/stable/3559865

Longford N.T. (2005), *Missing data and small area estimation,* Springer-Verlag, New York.

Longford N.T. (2006), *Missing data and small-area estimation: Modern analytical equipment for the survey statistician,* Springer Science & Business Media, Heidelberg-Berlin.

Lopez–Vizcaino E., Lombardia M.J., Morales D. (2015), *Small area estimation of labour force indicators under a multinomial model with correlated time and area effects*, "Journal of the Royal Statistical Society, Series A (Statistics in Society)", 178, 3, 535–565, https://www.jstor.org/stable/43965751

Łaciak J. (2013), *Aktywność turystyczna mieszkańców Polski w wyjazdach turystycznych w 2012 roku*, Instytut Turystyki, Warszawa, https://www.msit.gov.pl/download/1/4172/wPolski20124e5c.pdf

Marchetti S., Secondi L. (2017), *Estimates of household consumption expenditure at provincial level in Italy by using small area estimation methods: "Real" comparisons using purchasing power parities*, "Social Indicators Research", 131, 1, pp. 215–234, https://doi.org/10.1007/s11205-016-1230-8

Markowicz I. (2012), *Statystyczna analiza żywotności firm*, Rozprawy i Studia, 835, Wydawnictwo Uniwersytetu Szczecińskiego, Szczecin.

Markowicz I. (2016), *Analiza trwania firm w woj. zachodniopomorskim*, "Wiadomości Statystyczne", 1, 44–61, http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-16b9d823-a359-4e7a-aca9-326a698b5e7b?printView=true

Matei A., Tille Y. (2005), *Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size*, "Journal of Official Statistics", 21, 4, 543–570, https://www.semanticscholar.org/paper/Evaluation-of-variance-approximations-and-in-with-Matei-Till%C3%A9/d5f6f227b980dd01f25733de850195e2ee9514ae

Mauro F., Molina I., Garcia-Abril A., Valbuena R., Ayuga-Tellez E. (2016), *Remote sensing estimates and measures of uncertainty for forest variables at different aggregation levels*, "Environmetrics", 27, 4, 225–238, https://doi.org/10.1002/env.2387

Mazurek-Łopacińska K. (2002), *Komunikowanie się przedsiębiorstw z rynkiem w latach 1995-2000 – w świetle wyników badań ankietowych*, "Prace Naukowe Akademii Ekonomicznej we Wrocławiu. Zarządzanie i Marketing", 19, 926, 19–37.

McAllister R.J., Goe S.J., Butler, E.W. (1973), *Tracking respondents in longitudinal surveys: Preliminary considerations*, "Public Opinion Quarterly", 37, 3, 413–416, https://doi.org/10.1086/268103

McCullagh P., Nelder J.A. (1989), *Generalized linear models*, Second ed., Chapman and Hall, London.

McCulloch Ch.E. (2003), *Generalized linear mixed models*, "NSF-CBMS Regional Conference Series in Probability and Statistics", 7, 1–84, https://www.jstor.org/stable/4153190

Mellow W. (1981), *Unionism and wages: A longitudinal analysis*, "The Review of Economics and Statistics", LXII, 1, 43–52.

Menec V., Lix L., Steinbach C., Ekuma O., Sirski M., Dahl M., Soodeen R.A. (2004), *Patterns of health care use and cost at the end of life*, Manitoba Centre for Health Policy, Winnipeg, MB.

Menegaki A.N. (2011), *Growth and renewable energy in Europe: A random effect model with evidence for neutrality hypothesis*, "Energy Economics", 33, 2, 257–263, https://doi.org/10.1016/j.eneco.2010.10.004

Mikulec A. (2017), *Kohortowe tablice trwania przedsiębiorstw w województwie łódzkim – ujęcie kwartalne*, "Taksonomia 28. Klasyfikacja i analiza danych – teoria i zastosowania", 148–160, https://doi.org/10.15611/pn.2017.468.15

Militino A.F., Ugarte M.D., Goicoa T., Gonzalez-Audicana M. (2006), *Using small area models to estimate the total area occupied by olive trees*, "Journal of Agricultural, Biological, and Environmental Statistics", 11, 4, 450–461, https://doi.org/10.1198/108571106X154650

Mincer J. (1981), *Union effects: Wages, turnover, and job training*, "Working Paper", 808, National Bureau of Economic Research, Cambridge, MA, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=351352

Molina I. (2009), Un*certainty under a multivariate nested-error regression model with logarithmic transformation*, "Journal of Multivariate Analysis", 100, 963–980.

Molina I., Martin N. (2018), *EBP under a nested error model with log transformation*, "Annals of Statistics", 46, 5, 1961–1993.

Molina I., Rao J.N.K. (2010), *Small area estimation of poverty indicators*, "Canadian Journal of Statistics", 38, 3, 369–385, https://www.jstor.org/stable/27896031

Molina I., Saei A., Lombardia M.J. (2007), *Small area estimates of labour force participation under a multinomial logit mixed model*, "Journal of the Royal Statistical Society, Series A (Statistics in Society)", 170, 4, 975–1000, https://www.jstor.org/stable/4623223

Molina I., Salvati N., Pratesi M. (2008), *Bootstrap estimation of the mean squared error under a Spatial Fay-Herriot model*, "Comput Stat", 24, 441–458, https://doi.org/10.1007/s00180-008-0138-4

Molina I., Salvati N., Pratesi M. (2009), *Bootstrap for Estimating the MSE of the Spatial EBLUP*, "Computational Statistics", 24, 441–458, https://link.springer.com/article/10.1007/s00180-008-0138-4

Morales D., Esteban M.D., Perez A., Hobza T. (2021), *A course on small area estimation and mixed models methods. Theory and applications in R*, Springer, Cham, Switzerland.

Morris C.N., Christiansen C.L. (1996), *Hierarchical models for ranking and for identifying extremes with applications* [in:] *Bayes statistics*, J. M. Bernardo, J. O. Berger, A.P. Dawid, A.F.M. Smith (Eds.), 5, 277–297,Oxford University Press, Oxford.

Mukhopadhyay P. (1998), *Small area estimation in survey sampling*, Narosa Publishing House, New Delhi.

Munnell A.H. (1990), *Why has productivity growth declined? Productivity and public investment*, “New England Economic Review”, 3–22, https://www.bostonfed.org/-/media /Documents/neer/neer190a.pdf

Nair K.R. (1941), *A note on the method of ‘fitting of constants’ for analysis of non-orthogonal data arranged in a double classification*, “Sankhyā: The Indian Journal of Statistics”, 5, 3, 317–328, https://www.jstor.org/stable/25047696

Namazi-Rad M.-R., Steel D. (2015), *What level of statistical model should we use in small area estimation?* “Australian & New Zealand Journal of Statistics”, 57, 2, 275–298, https://doi.org/10.1111/anzs.12115

Nathan G. (2009), *The analysis of longitudinal surveys* [in:] *Sample surveys: Design, methods and applications*, D. Pfefformann, C. R. Rao (Eds.), Handbook of Statistics, 29B, 315–327, Elsevier, Amsterdam.

Nathan G., Holt D. (1980), *The effect of survey design on regression analysis*, “Journal of the Royal Statistical Society, Series B (Methodological)”, 42, 3, 377–386, https://www.jstor.org/stable/2985175

National Health Interview Survey (2019), *NHIS sample questionnaire brochure*, National Center for Health Statistics, https://www.cdc.gov/nchs/data/nhis/sample-questionnaire-brochure.pdf

National Research Council (U.S.) (1980), *Issues and current studies*, National Academy of Sciences, 190, Washington, DC.

Nekrasaite-Liege V., Radavicius M., Rudys T. (2011), *Model-based design in small area estimation*, “Lithuanian Mathematical Journal”, 51, 3, 417–424, https://doi.org/10.1007/s10986-011-9136-2

Nerlove M. (2002), *Essays in panel data econometrics*, Cambridge University Press, Cambridge.

Neter J., Waksberg J. (1964), *A study of response errors in expenditures data from household interviews*, “Journal of the American Statistical Association”, 59, 305, 18–55, https://doi.org/10.1080/01621459.1964.10480699

Niemiro W., Wesołowski J. (2010), *Synthetic and composite estimation under a superpopulation model*, “Statistical Papers”, 51, 3, 497–509, https://doi.org/10.1007/s00362-008-0136-1

Nusser S.M., Goebel J.J. (1997), *The National Resources Inventory: A long-term multi-resource monitoring programme*, “Environmental and Ecological Statistics”, 4, 181–204, https://doi.org/10.1023/A:1018574412308

Ogungbenro K., Graham G., Gueirguieva I., Aarons L. (2008), *Incorporating correlation in interindividual variability for the optimal design of multiresponse pharmacokinetic experiments*, "Journal of Biopharmaceutical Statistics", 18, 342–358, https://doi.org/10.1080/10543400701697208

Opsomer J.D., Claeskens G., Ranalli M.G., Kauermann G., Breidt F.J. (2008), *Non-parametric small area estimation using penalized spline regression*, "Journal of the Royal Statistical Society, Series B (Statistical Methodology)", 70, 1, 265–286, https://www.jstor.org/stable/20203822

Ott N. (1995), *The use of panel data in the analysis of household structures* [in:] *Household demography and household modeling*, E. Imhoff, A. Kuijsten, P. Hooimeijer, L. Wissen (Eds.), Springer, Boston, MA, 163–183, https://doi.org/10.1007/978-1-4757-5424-7_7

Paradysz J. (1998), *Small area statistic in Poland – first experiences and application possibilities*, "Statistics in Transition", 3, 5, 1003–1015.

Pawłowski Z. (1969), *Ekonometria*, PWN, Warszawa.

Pawłowski Z. (1981), *Elementy ekonometrii*, PWN, Warszawa.

Pawlowsky-Glahn V., Buccianti A., Eds. (2011), *Compositional data analysis*, Wiley, Chichester.

Pedone P., Romano D. (2011), *Designing small samples for form error estimation with coordinate measuring machines*, "Precision Engineering", 35, 2, 262–270, https://doi.org/10.1016/j.precisioneng.2010.10.002

Pereira L.N., Coelho P.S. (2013), *Estimation of house prices in regions with small sample sizes*, "The Annals of Regional Science", 50, 2, 603–621, https://doi.org/10.1007/s00168-012-0507-3

Pesaran M.H. (2003), *On aggregation of linear dynamic models: An application to life-cycle consumption models under habit formation*, "Economic Modeling", 20, 227–435, https://doi.org/10.1016/S0264-9993(02)00059-7

Petrucci A., Pratesi M., Salvati N. (2005), *Geographic information in small-area estimation: Small-area models and spatially correlated random area effects*, "Statistics in Transition", 7, 609–623, https://www.researchgate.net/publication/228758984

Petrucci A., Salvati N. (2004), *Small area estimation considering spatially correlated errors: The unit level random effects model*, University of Florence, Florence, https://local.disia.unifi.it/pubblicazioni_DS/wp/2004/wp2004_10.pdf

Petrucci A., Salvati N. (2006), *Small area estimation for spatial correlation in watershed erosion assessment*, "Journal of Agricultural, Biological and Environmental Statistics", 11, 2, 169–182, https://www.jstor.org/stable/27595594

Pfeffermann D. (1993), *The role of sampling weights when modeling survey data*, "International Statistical Review", 317–337, https://doi.org/10.2307/1403631

Pfeffermann D., Correa S. (2012), *Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation*, "Biometrika", 99, 2, 457–472, https://www.jstor.org/stable/41720703

Pfeffermann D., Glickman H. (2004), *Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap*, "Proceedings of the Survey Research Methods Section American Statistical Association", 4167–4178, https://www.researchgate.net/publication/252939945

Pfeffermann D., Moura F., Silva P.N. (2001), *Multi-level modeling under informative probability sampling*, S3RI Methodology Working Papers, M04/09, Southampton Statistical Sciences Research Institute, Southampton, https://eprints.soton.ac.uk/8182/1/8182-01.pdf

Pfeffermann D., Tiller R. (2006), *Small-area estimation with state-space models subject to benchmark constraints*, "Journal of the American Statistical Association", 101, 476, 1387–1397, https://www.jstor.org/stable/27639759

Phillips P.C., Durlauf S.N. (1986), *Multiple time series regression with integrated processes*, "The Review of Economic Studies", 53, 4, 473–495, https://doi.org/10.2307/2297602

Piekałkiewicz J. (1934), *Sprawozdanie z badań składu ludności robotniczej w Polsce metodą reprezentacyjną: na podstawie materiałów spisu powszechnego ludności w d. 9. XII. 1931*, Instytut Spraw Społecznych, Warszawa, http://bc.gbpizs.gov.pl/dlibra/publication/706/edition/680/content

Pietrzak M.B. (2010), *Wykorzystanie odległości ekonomicznej w przestrzennej analizie stopy bezrobocia dla Polski*, "Oeconomia Copernicana", 1, 79–98, http://bazekon.icm.edu.pl/bazekon/element/bwmeta1.element.ekon-element-000171257333

Pinheiro J.C., Bates D.M. (1995), *Model building for nonlinear mixed-effects models*, Technical Report, 91, Department of Biostatistics, University of Wisconsin Madison, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=819350aa1cafde32d7739ad78e1e50781c8fc7a7

Pinheiro J.C., Bates D.M. (2000), *Mixed-effects models in S and S-Plus*, Springer, New York.

Płatek R., Rao J.N.K., Sarndal C.E., Singh M.P., Eds. (1987), *Small area statistics: An international symposium*, John Wiley & Sons, New York, https://doi.org/10.2307/2348315

Prasad N.G.N., Rao J.N.K. (1990), *The estimation of the mean squared error of small-area estimators*, "Journal of the American Statistical Association", 85, 409, 163–171, https://doi.org/10.2307/2289539

Prasad N.G.N., Rao J.N.K. (1999), *On robust small area estimation using a simple random effects model*, Catalogue No. 12-001-X, Business Survey Methods Division, Statistics Canada, 67–72, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1999001/article/4713-eng.pdf?st=E7CNaMLx

Pratesi M. (2015), *Spatial disaggregation and small-area estimation methods for agricultural surveys: Solutions and perspectives*, Technical Report Series GO-07-2015, Global Strategy to Improve Agricultural and Rural Statistics, World Bank, Washington, DC.

Pratesi M., Ed. (2016), *Analysis of poverty data by small area estimation*, John Wiley & Sons, Hoboken, NJ.

Pratesi M., Salvati N. (2008), *Small area estimation: The EBLUP estimator based on spatially correlated random area effects*, "Statistical Methods and Applications", 17, 1, 113–141, https://doi.org/10.1007/s10260-007-0061-9

Pratesi M., Salvati N. (2009), *Small-area estimation in the presence of correlated random area effects*, "Journal of Official Statistics", 25, 1, 37–53, https://www.researchgate.net/publi cation/281159905

Purcell N.J., Kish L. (1979), *A biometrics invited paper. Estimation for small domains*, "Biometrics", 35, 365–384, https://doi.org/10.2307/2530340

Purcell N.J., Kish L. (1980), *Postcensal estimates for local areas (or domains)*, "International Statistical Review", 48, 3–18, https://doi.org/10.2307/1402400

R Core Team (2022), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna.

Raghunath A. (1990), *On comutativity of design and model expectations in randomized response surveys*, "Communications in Statistics-Theory and Methods", 19, 10, 3751–3757, https://doi.org/10.1080/03610929008830411

Rahman A., Harding A. (2016), *Small area estimation and microsimulation modeling*, CRC Press, Boca Raton, FL.

Rao J.N.K. (2003), *Small area estimation*, John Wiley & Sons, New York, https://doi.org/10.1002/0471722189

Rao J.N.K., Molina I. (2015), *Small area estimation*, John Wiley & Sons, Hoboken, NY, https://doi.org/10.1002/9781118735855

Rao J.N.K., Sinha S.K., Dumitrescu L. (2014), *Robust small area estimation under semi-parametric mixed models*, "The Canadian Journal of Statistics", 42, 1, 126–141, https://www.jstor.org/stable/43185172

Rao J.N.K., Yu M. (1992), *Small area estimation by combining time series and cross-sectional data*, "Proceedings of the Section on Survey Research Method, American Statistical Association", 1–9.

Rao J.N.K., Yu M. (1994), *Small-area estimation by combining time-series and cross-sectional data*, "Canadian Journal of Statistics", 22, 4, 511–528, https://doi.org/10.2307/3315407

Rao T.V.H. (1962), *An existence theorem in sampling theory*, "Sankhyā, Series A", 24, 327–330.

Reinsel G. (1984), *Estimation and prediction in a multivariate random effects generalized linear model*, "Journal of the American Statistical Association", 79, 386, 406–414, https://doi.org/10.2307/2288283

Ręklewski M., Śliwicki D. (2016), *Estymacja dla małych obszarów liczby biernych* zawodowo *w powiatach woj. kujawsko-pomorskiego*, "Wiadomości Statystyczne", 5, 37–47.

Richardson A.M. (1997), *Bounded influence estimation in the mixed linear model*, "Journal of the American Statistical Association", 92, 437, 154–161, https://doi.org/10.2307/2291459

Richardson A.M., Welsh A.H. (1995), *Robust restricted maximum likelihood in mixed linear models*, "Biometrics", 51, 1429–1439, https://doi.org/10.2307/2533273

Rivest L.P., Verret F., Baillargeon S. (2016), *Unit level small area estimation with copulas*, "Canadian Journal of Statistics", 44, 4, 397–415, https://www.jstor.org/stable/44708498

Robinson G.K. (1991), *That BLUP is a good thing: The estimation of random effects*, "Statistical Science", 6, 15–31, https://doi.org/10.1214/ss/1177011926

Royall R.M. (1976), *The linear least squares prediction approach to two-stage sampling*, "Journal of the American Statistical Association", 71, 657–473, https://doi.org/10.2307/2285596

Royall R.M., Cumberland W.G. (1981), *An empirical study of the ratio estimator and estimators of its variance*, "Journal of the American Statistical Association", 76, 66–88, https://doi.org/10.1080/01621459.1981.10477604

Rueda C., Menendez J.A., Gomez F. (2010), *Small area estimators based on restricted mixed models*, "TEST", 19, 558–579, https://doi.org/10.1007/s11749-010-0186-2

Saei A., Chambers R. (2003), *Small area estimation under linear and generalized linear mixed models with time and area effects*, S3RI Methodology Working Paper M03/15r, University of Southampton, Southampton, UK.

Salvati N., Chandra H., Chambers R. (2012a), *Model-based direct estimation of small-area distributions*, "Australian & New Zealand Journal of Statistics", 54, 1, 103–123.

Salvati N., Tzavidis N., Pratesi M., Chambers R. (2012b), *Small area estimation via M-quantile geographically weighted regression*, "TEST", 21, 1–28, https://doi.org/10.1007/s11749-010-0231-1

Sarndal C.-E. (1981), *Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse*, "Bulletin of the International Statistical Institute", 49, 494–513.

Sarndal C.-E. (2010), *Models in survey sampling*, "Statistics in Transition – New Series", 3, 11, 112–127, https://sit.stat.gov.pl/SiT/2010/3/7_Sarndal112-127.pdf

Sarndal C.-E., Hidiroglou M.A. (1989), *Small domain estimation: A conditional analysis*, "Journal of the American Statistical Association", 84, 266–275, https://doi.org/10.2307/2289873

Sarndal C.-E., Swensson B., Wretman J. (1992), *Model assisted survey sampling*, Springer-Verlag, New York.

Schaible W.L. (1993), *Use of small area estimators in U.S. federal programs [in:] Small area statistics and survey designs*, International Scientific Conference, vol. I. Invited papers, GUS, Warsaw, 95–114.

Schmid T., Munnich R.T. (2014), *Spatial robust small area estimation*, "Statistical Papers", 55, 653–670, https://doi.org/10.1007/s00362-013-0517-y

Schmid T., Tzavidis N., Munnich R., Chambers R. (2016), *Outlier robust small area estimation under spatial correlation*, "Scandinavian Journal of Statistics", 43, 806–826.

Schwartz G. (1978), *Estimating the dimension of a model*, "Annals of Statistics", 6, 2, 461–464.

Sen A.R. (1953), *On the estimate of the variance in sampling with varying probabilities*, "Journal of the Indian Society of Agricultural Statistics", 5, 119–127, https://link.springer.com/content/pdf/10.1007/BF02868860.pdf

Shapiro S.S., Wilk M.B. (1965), *An analysis of variance test for normality (complete samples)*, "Biometrika", 52, 3/4, 591–611, https://doi.org/10.2307/2333709

Sharot T. (1991), *Attrition and rotation in panel surveys*, "The Statistician", 325–331, https://doi.org/10.2307/2348285

Singh B., Shukla G., Kundu D. (2005), *Spatio-temporal models in small area estimation*, "Survey Methodology", 31, 2, 183–195, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9053-eng.pdf?st=r_r6oam0

Sinha S.K., Rao J.N.K. (2009), *Robust small area estimation*, "The Canadian Journal of Statistics", 37, 3, 381–399, https://www.jstor.org/stable/25653486

Sinha S.K., Sattar A. (2015), *Inference in semi-parametric spline mixed models for longitudinal data*, "Metron", 73, 3, 377–395, https://doi.org/10.1007/s40300-015-0059-2

Siswantining T., Naima M.G., Soemartojo S.M. (2020), *Estimation of variance of random effect in small area model with Spatial Empirical Best Linear Unbiased Prediction (SEBLUP)*, "Journal of Physics: Conference Series", 1442, 1, 12–32, https://doi.org/10.1088/1742-6596/1442/1/012032

Słomczyński K.M. with the team (2014), *POLPAN 1988-2013. Podstawowe informacje o Polskim Badaniu Panelowym*, Zespół Porównawczych Analiz Nierówności Społecznych, Instytut Filozofii i Socjologii Polskiej Akademii Nauk, Warszawa 2014.

Smith T.M.F. (1994), *Sample survey, 1975-1990 an age of reconciliation?* "International Statistical Review", 62, 3–34, https://doi.org/10.2307/1403539

Sobczyk M. (2004), *Statystyka*, Wydawnictwo Naukowe PWN, Warszawa.

Sobczyk M. (2012), *Statystyka*, Wydawnictwo Naukowe PWN, Warszawa.

Spława-Neyman J. (1933), *Zarys teorii i praktyki badania struktury ludności metoda reprezentacyjna*, Instytut Spraw Społecznych, Warszawa.

Stachurski T. (2018), *A simulation analysis of the accuracy of median estimators for different sampling designs* [in:] *36th International Conference Mathematical Methods in Economics. MME 2018*, Conference proceedings, L. Vachova, V. Krotochvil (Eds.), Matfyz Press, Praha, 509–514, https://www.researchgate.net/publication/364639125_A_simulation_analysis_of_the_accuracy_of_median_estimators_for_different_sampling_designs

Stahel W.A., Welsh A.H. (1992), *Robust estimation of variance components*, Technical Report 69, Seminar fur Statistik, ETH Zurich.

Starzyńska W. (2005), *Statystyka praktyczna*, Wydawnictwo Naukowe PWN, Warszawa.

Statistics Poland (2011), *Metodologia badania budżetów gospodarstw domowych*, Warszawa.

Statistics Poland (2012), *Warunki powstania i działania oraz perspektywy rozwojowe polskich przedsiębiorstw powstałych w latach 2006–2010*, Warszawa.

Statistics Poland (2013), *Aktywność ekonomiczna ludności Polski, II kwartał 2013*, Warszawa.

Statistics Poland (2015), *Warunki powstania i działania oraz perspektywy rozwojowe polskich przedsiębiorstw powstałych w latach 2009–2013*, Warszawa.

Statistics Poland (2017a), *Budżety gospodarstw domowych w 2016 r.*, Warszawa.

Statistics Poland (2017b), *Aktywność ekonomiczna ludności Polski, II kwartał 2017*, Warszawa.

Steel D.G. (2004), *Sampling in time* [in:] *Encyclopedia of Social Measurement*, Kimberly Kempf-Leonard (Ed.), Academic Press, San Diego, CA, 823–828.

Steel D., McLaren C. (2009), *Design and analysis of surveys repeated over time* [in:] *Sample surveys: Inference and analysis*, C.R. Rao (Ed.), Handbook of Statistics, 29, Elsevier, Amsterdam, 289–313.

Strahl D., Sobczak E., Markowska M., Bal-Domańska B. (2004), *Modelowanie ekonometryczne z Excelem. Materiały pomocnicze do laboratorium z ekonometrii*, Wydawnictwo Akademii Ekonomicznej, Wrocław.

Stram D.O., Lee J.W. (1994), *Variance components testing in the longitudinal mixed effects model*, "Biometrics", 1171–1177, https://doi.org/10.2307/2533455

Stukel D.M., Rao J.N.K. (1999), *On small-area estimation under two-fold nested error regression models*, "Journal of Statistical Planning and Inference", 78, 1–2, 131–147, https://doi.org/10.1016/S0378-3758(98)00211-0

Suchecki B., Ed. (2010), *Ekonometria przestrzenna. Metody i modele analizy danych przestrzennych*, Wydawnictwo C.H. Beck, Warszawa.

Sud U.C., Bhatia V.K., Chandra H., Srivastava A.K. (2011), *Crop yield estimation at district level by combining improvement of crop statistics scheme data and census data*, Wye City Group on Rural Statistics and Agricultural Household Income, 4th Meeting, Rio de Janeiro, https://www.fao.org/fileadmin/templates/ess/pages/rural/wye_city_group/2011/documents/Session9/Sud__Bhatia__Chandra__Srivastava_-_Paper.pdf

Sudman S., Ferber R. (1970), *Consumer panels*, American Marketing Association, Chicago.

Sztumski J. (2004), *Metoda monograficzna, jej zalety i niedostatki*, "Zeszyty Naukowe Górnośląskiej Wyższej Szkoły Handlowej", 25, 7–16.

Szymkowiak M. (2020), *Podejście kalibracyjne w badaniach społeczno-ekonomicznych*, Wydawnictwo Uniwersytetu Ekonomicznego, Poznań.

Tanton R., Vidyattama Y., Nepal B., McNamara J. (2011), *Small area estimation using a reweighting algorithm*, "Journal of the Royal Statistical Society, Series A (Statistics in Society)", 174, 4, 931–951, https://doi.org/10.1111/j.1467-985X.2011.00690.x

Tanur J.M. (1981), *Advances in methods for large-scale surveys and experiments* [in:] *The 5-year outlook on science and technology 1981: Source materials*, Vol. 2, National Science Foundation, Washington, DC.

Taylor M.F., Brice J., Buck N., Prentice-Lane E., Eds. (2018), *British household panel survey, User manual*, Vol. A: *Introduction, technical report and appendices*, University of Essex, Colchester, http://doi.org/10.5255/UKDA-SN-5151-2

Thompson W.A., Jr. (1962), *The problem of negative estimates of variance components*, "Annals of Mathematical Statistics", 33, 273–289, https://doi.org/10.1214/aoms/1177704731

Tille Y. (2006), *Sampling algorithms*, Springer, New York.

Torabi M., Rao J.N.K. (2010), *Mean squared error estimators of small area means using survey weights*, “Canadian Journal of Statistics”, 38, 4, 598–608, https://doi.org/10.1002/cjs.10078

Torabi M., Shokoohi F. (2015), *Non-parametric generalized linear mixed models in small area estimation*, “Canadian Journal of Statistics”, 43, 1, 82–96, https://doi.org/10.1002/cjs.11236

Trivellato U. (1999), *Issues in the design and analysis of panel studies: A cursory review*, “Quality and Quantity”, 33, 3, 339–351, https://doi.org/10.1023/A:1004657006031

Ugarte M.D., Goicoa T., Militino A.F., Durban M. (2009), *Spline smoothing in small area trend estimation and forecasting*, “Computational Statistics & Data Analysis”, 53, 3616–3629, https://doi.org/10.1016/j.csda.2009.02.027

Valliant R., Dorfman A.H., Royall R.M. (2000), *Finite population sampling and inference: A prediction approach*, Wiley Series in Probaility and Statistics, John Wiley & Sons, New York.

Verbeke G., Molenberghs G. (2000), *Linear mixed models for longitudinal data*, Springer-Verlag, New York.

Vonesh E.F., Carter R.L. (1992), *Mixed-effects nonlinear regression for unbalanced repeated measures,* “Biometrics”, 1–17, https://doi.org/10.2307/2532734

Wang J., Fuller W.A. (2003), *The mean squared error of small area predictors constructed with estimated area variances*, “Journal of the American Statistical Association”, 98, 463, 716–723, https://www.jstor.org/stable/30045299

Wansbeek T.J., Koning R.H. (1989), *Measurement error and panel data*, “Statistica Neerlandica”, 45, 85–92, https://doi.org/10.1111/j.1467-9574.1991.tb01296.x

Watson D.J. (1937), *The estimation of leaf area in field crops*, “The Journal of Agricultural Science”, 27, 3, 474–483, https://doi.org/10.1017/S002185960005173X

Wawrowski Ł. (2012), *Analiza ubóstwa w przekroju powiatów w województwie wielkopolskim z wykorzystaniem metod statystyki małych obszarów*, “Przegląd Statystyczny”, 59, 248–260, https://bazekon.uek.krakow.pl/rekord/171236135

Witte I. (1987), *Haushalt und Familie* [in:] *Statistisches Bundesamt, Datenrepon 1987 – Zahlen und Fakten über die BRD*, Deutschland Statistisches Bundesamt, Bundeszentrale für politische Bildung, 257, Bonn, 368–376.

Witte I. (1988), *Haushalt und Familie* [in:] *Datenband 1987: Lebenslagen im Wandel,* H.J. Krupp, J. Schupp (Eds.), Campus Verlag, Frankfurt, 21–41.

Witte I., Lahmann H. (1988), *Residential mobility of one-person households* [in:] *Fourth Annual Research Conference – Proceedings, Bureau of the Census*, Washington D.C., 422–448.

Wolfinger R. (1993), *Covariance structure selection in general mixed models*, "Communications in Statistics – Simulation and Computation", 22, 4, 1079–1106, https://doi.org/10.1080/03610919308813143

Wright T. (2001), *Selected moments in the development of probability sampling: Theory & practice*, "Survey Research Methods Section Newsletter U.S.", Census Bureau, 13, 1–6.

Wu C.-F. (1982), *Estimation of variance of the ratio estimator*, "Biometrika", 69, 1, 183–189, https://doi.org/10.2307/2335867

Wywiał J.L. (2010), *Wprowadzenie do metody reprezentacyjnej*, Wydawnictwo Akademii Ekonomicznej, Katowice.

Yates F., Grundy P.M. (1953), *Selection without replacement from within strata with probability proportional to size*, "Journal of the Royal Statistical Society", Series B, 13, 253–261, https://www.jstor.org/stable/2983772

You Y., Rao J.N.K. (2002), *A pseudo-empirical best linear unbiased prediction aproach to small area estimation using survey weights*, "The Canadian Journal of Statistics", 30, 3, 431–439, https://doi.org/10.2307/3316146

You J., Datta G.S., Maples J.J. (2014), *Modeling disability in small areas: An area-level approach of combining two surveys*, "Statistics", 11, https://www.census.gov/content /dam/Census/library/working-papers/2014/adrm/rrs2014-11.pdf

Zeliaś A., Pawełek B., Wanat S. (2002), *Metody statystyczne: zadania i sprawdziany*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Żądło T. (2004), *On unbiasedness of some EBLU predictor* [in:] *Proceedings in computational statistics 2004*, J. Antoch (Ed.), Physica Verlag, Heidelberg-New York, 2019–2026.

Żądło T. (2008), *Elementy statystyki małych obszarów z programem R*, Wydawnictwo Akademii Ekonomicznej, Katowice.

Żądło T. (2009), *On MSE of EBLUP*, "Statistical Papers", 50, 101–118, https://doi.org// 10.1007/s00362-007-0066-3

Żądło T. (2015), *Statystyka małych obszarów w badaniach ekonomicznych. Podejście modelowe i mieszane*, Wydawnictwo Uniwersytetu Ekonomicznego, Katowice.

Żądło T. (2017), *On prediction of population and subpopulation characteristics for future periods*, "Communications in Statistics – Simulation and Computation", 46, 10, 8086–8104, https://doi.org/10.1080/03610918.2016.1263737

Żądło T. (2020), *On accuracy estimation using parametric bootstrap in small area prediction problems*, "Journal of Official Statistics", 36, 2, 435–458, https://doi.org/10.2478/jos-2020-0022

List of figures

List of tables

The growing importance of regions and regional policy (…) also entails an increase in the importance of national databases with a very detailed territorial division – sources increasingly used by public statistics (…) This is also related to the growing demand for information at an increasingly lower level of aggregation, as well as the demand for methods that do not require large financial outlays, but make it possible to obtain accurate estimations of subpopulation characteristics quickly, without the need for a full survey. Small area estimation methods may be the answer to this demand, allowing estimation and prediction under conditions where classical estimation methods prove to be inefficient or too costly. They allow estimation even for very small sample sizes, and even when the sample size of a subpopulation is zero. The choice of the topic considered in this monograph is therefore related to the increasing demand for local cross-sectional analyses. Moreover, it is also due to the multitude of fields in which the methods of small area estimation have already found application, such as market analyses, regional policy, labour market and poverty analysis, agricultural economics, and economic aspects of health policy.

Excerpt from the book

PhD Małgorzata Krzciuk – Assistant Professor in the Department of Statistics, Econometrics and Mathematics at the University of Economics in Katowice. Winner of the Competition of the President of the Statistics Poland for the best doctoral thesis on statistics in 2022. The author's research interests focus on issues of small area estimation, survey sampling, multivariate statistical analysis and computer simulation techniques. Author of numerous publications in Polish and English and participant in many national and international scientific conferences including the II Congress of Polish Statistics, the 6th ITAlian COnference on Survey Methodology ITACOSM 2019 and SAE BIG4small the Satellite Conference of the 63rd ISI World Statistics Congress. Author conducts laboratories of subjects such as Small area estimation, Sampling in Economic Analysis, Statistical Inference, and Longitudinal Data Analysis. Member of the Polish Statistical Association.

ISBN 978-83-7875-886-0

University
of Economics
in Katowice