The interplay between sampling design and statistical modelling in small area estimation

Submitted in partial fulfilment of the requirements for the degree

Dr. rer. pol.

to the
Department IV
at the University of Trier
submitted by

Diplom Volkswirt Thomas Zimmermann
Kloschinskystraße 59, 54292 Trier
born 21.02.1985 in Cologne

Supervisors:
Prof. Dr. Ralf Münnich (Universität Trier)
Prof. Dr. Li-Chun Zhang (University of Southampton)

Contents

Acknowledgement

First and foremost, I would like to express my deep and sincere gratitude to my supervisor, Professor Dr. Ralf Münnich. His invaluable support, personal guidance and trust in me have a provided a very fruitful basis for this thesis.

I am also deeply grateful to Professor Dr. Li-Chun Zhang for his detailed and constructive comments as well as his important support.

I wish to express my thanks to the BLUE-ETS project, which provided funding for my position as research assistant at the University of Trier.

My thanks also go the IAOS for their award of the second prize in the young statisticians competition.

It is my pleasure to thank all my colleagues at the University of Trier for the extremely pleasant and supportive work atmosphere, in particular Dr. Jan Pablo Burgard, Dr. Florian Ertz and Dr. Jan Georg Seger.

My special gratitude is due to my brother and parents for their loving support.

Finally, I owe my loving thanks to my girlfriend Charlotte Kaplan. She has been a great source of patience, understanding and encouragement throughout all these years.

Zusammenfassung

Traditionell werden Stichprobenerhebungen so geplant, dass nationale Statistiken mit einer adäquaten Präzision geschätzt werden können. Dies kann zu sehr kleinen Stichprobenumfängen für bestimmte Subpopulationen führen, so dass direkte, designbasierte Schätzmethoden keine Schätzungen für besagte Untergruppen mit einer akzeptablen Genauigkeit erlauben. Einen Ausweg aus diesem Dilemma stellt die Verwendung modellbasierter Schätzverfahren dar, welche auch bei kleinen Stichprobenumfängen noch präzise Schätzungen erlauben. Eine Besonderheit der modellbasierten Verfahren ist, dass in vielen Fällen keinerlei Designinformationen bei der Schätzung betrachtet werden. Hieraus können Verzerrungen resultieren, welche die Anwendbarkeit besagter modellbasierter Verfahren stark einschränken.

Die vorliegende Arbeit beschäftigt sich daher speziell mit dem Zusammenspiel zwischen dem Stichprobendesign und statistischen Modellierungen im Bereich der Small Area – Statistik. Dabei werden insbesondere zwei Fragestellungen betrachtet:

1. Wenn wir bereits wissen, dass wir später statistische Modelle für die Stichprobendaten schätzen müssen, wie können wir dann ein Stichprobendesign so ausgestalten, dass nationale Statistiken präzise geschätzt werden können, gleichzeitig aber keine Verzerrungen für modellbasierte Schätzverfahren resultieren?

2. Wenn erst nach Ziehung der Stichprobe bekannt wird, dass modellbasierte Small Area – Schätzungen benötigt werden, wie können wir dabei das Stichprobendesign angemessen berücksichtigen, so dass Verzerrungen vermieden werden?

In dieser Arbeit werden nach einer Vorstellung des obigen Zielkonflikts designbasierte Schätzmethoden vorgestellt, die für große und geplante Domains ausreichend präzise Ergebnisse liefern. Dabei wird auch auf Ansätze der Stichprobenplanung für die Anwendung designbasierter Schätzverfahren im Small Area – Kontext eingegangen. Anschließend werden gängige modellbasierte Small Area – Schätzverfahren vorgestellt, wobei neben der Mittelwertschätzung aus gemischten linearen Modellen ein Schwerpunkt auf die Small Area – Schätzung unter nicht-linearen Transformationen gelegt wird. Schließlich werden verschiedene Ansätze zur Auswahl eines geeigneten statistischen Modells sowie zur Überprüfung der Modellannahmen diskutiert und anhand zweier Datensätze illustriert.

Im Folgenden wird das Problem der Verzerrungen modellbasierter Verfahren aufgrund des Stichprobendesigns ausführlich erörtert und verschiedene Lösungsstrategien für gemischte lineare Modelle präsentiert. Darauf aufbauend werden Vorschläge zur Vermeidung besagter Verzerrungen für den optimalen Schätzer unter einem lognormalverteilten gemischten Modell bei Unit Level – Informationen entwickelt. Dieses Problem wurde bislang in der Literatur noch vernachlässigt. Als Lösungsansatz wird in dieser Arbeit

ein optimaler Schätzer unter einem erweiterten Modell vorgeschlagen, wobei das Modell durch die Berücksichtigung einer Funktion des Design-Gewichts als zusätzlicher Kovariable ergänzt wird. Für diesen Schätzer werden anschließend Ansätze zur Schätzung des mittleren quadratischen Fehlers (MSE) herausgearbeitet. Die Ergebnisse einer Simulationsstudie demonstrieren die Eignung des vorgeschlagenen Schätzers zur verlässlichen Schätzung trotz Verzerrungen aufgrund des Stichprobendesigns. Desweiteren wird ein praktikabler Methodenmix präsentiert, der die Auswahl einer geeigneten Variable zur Modellerweiterung ermöglicht.

Anschließend wird ein neues Konzept zur Vermeidung von informativen Stichprobendesigns erarbeitet, welches trotzdem eine präzise Schätzung von nationalen Statistiken mittels designbasierter Verfahren erlaubt. Das Konzept verfolgt die Idee, entsprechend einer Hilfsvariablen die Einheiten der Population so zu klumpen, dass die Einheiten innerhalb eines Klumpens möglichst heterogen sind. Es resultiert ein Stichprobendesign, welches die Schätzung von Modellen nicht stört, und für eine Vielzahl von praxisrelevanten Situationen eine präzise Schätzung nationaler Statistiken erlaubt. Dies wird für einige Modelle theoretisch nachgewiesen. Darüber hinaus erfolgt ein Vergleich mit anderen Varianzreduktionsverfahren im Rahmen von Simulationsstudien. Dabei zeigt sich auch das große Potenzial der entwickelten Methode zur Kompensation einer etwaigen Modellfehlspezifikation sowie zur präziseren modellbasierten Schätzung von Armutsgefährdungsquoten, wenn die Armutsgrenze aus der Stichprobe geschätzt werden soll.

Schließlich werden in einem weiteren Kapitel ausgewählte Anwendungen von Small Area – Verfahren in einer designbasierten Umgebung mittels Simulationsstudien präsentiert. Die erste Anwendung bezieht sich auf die Small Area – Schätzung für Unternehmensstichproben. Hierbei stellt sich vor allem die Problematik extrem schiefer Verteilungen, so dass die Anwendbarkeit von Modellen sehr erschwert wird. Hiernach folgt die Schätzung von Beschäftigten- und Arbeitslosenzahlen anhand der Luxemburger Arbeitskräfteerhebung. In diesem Fall ist die Verfügbarkeit von guten Hilfsinformationen stark eingeschränkt, so dass insbesondere die Frage im Raum steht, welche Modellierung bei modellassistierten Verfahren geeignet wird. Zum Schluss wird noch eine Studie zur Schätzung der Armutsgefährdungsquoten für Small Areas präsentiert. Hier wird neben der Frage, wie das Stichprobendesign aussehen könnte, insbesondere thematisiert, welche Art von Small Area – Modellierungen besonders aussichtsreich ist.

List of Figures

List of Tables

List of Algorithms

List of Symbols

$\mathbf{a}, \mathbf{A}$	Arbitrary vector and matrix
α	Arbitrary constant
$\mathbf{b}, \mathbf{B}$	Arbitrary vector and matrix
b_*	Size measure for unit $*$
$\boldsymbol{\beta}$	Vector of fixed-effects model parameters of length p including the intercept
$\mathrm{Bern}(\theta)$	Bernoulli distribution with success probability θ
$\mathrm{Bin}(n, \theta)$	Binomial distribution with success probability θ and n trials
$\mathbf{C}$	$n \times q$ incidence matrix for the random effects.
c	Arbitrary constant
$\mathrm{CV}(*)$	Coefficient of variation of an estimate $*$
CV_*	Upper boundary on the coefficient of variation
$\mathrm{Cov}(*, **)$	Covariance between $*$ and $**$
d, D	Running index $d = 1, \ldots, D$ with D as the total number of areas
$\mathbf{D}$	Variance-covariance matrix of the random effects
DEFF	Design effect
$\boldsymbol{\delta}$	Vector of variance components
e_*	Residual of unit $*$
E_*	Hypothetical census-fit residual of unit $*$
$\mathrm{E}(*)$	Expectation of $*$
$\boldsymbol{\varepsilon}$	$n \times 1$ vector of the sampling error
$\boldsymbol{\eta}$	Vector of mixed effects
$f(\cdot)$	Arbitrary function
f_*	Sampling fraction in area $*$
G	Relative priority for the national mean in the Longford allocation
$g(\cdot), G(\cdot)$	Arbitrary functions
g_{1*}, g_{2*}, g_{3*}	Components of the Prasad-Rao like MSE estimators in area $*$
γ_*	Shrinkage coefficient for area $*$
$h(\cdot)$	Arbitrary function
h	Running index for the PSUs
i, I	Running index $i = 1, \ldots, I$ with I as the total number of categories
$\mathbf{I}_*$	Identity matrix of dimension $* \times *$
$\mathbb{I}_*$	Indicator function, if $*$ is true, then $\mathbb{I}_* := 1$ else $\mathbb{I}_* := 0$
I_*	Sample membership indicator for unit $*$
ICC	Intraclass correlation coefficient
j	Running index for observations within an area $*$, $j = 1, \ldots, n_*$
K	Penalty term for model complexity in information criteria
k_*	Inverse sampling fraction in stratum $*$
$\boldsymbol{\kappa}$	Vector of regression coefficients for augmentation variables

L	The number of PSUs
L_*	Set of PSU indices
$\mathbf{l}$	Arbitrary vector
$\boldsymbol{\lambda}$	Vector of Lagrange multipliers
λ_*	Scaling factor for the weights in area $*$
$\bar{l}_*$	Sample mean of the log of the dependent variable in area $*$
$\text{logit}(\theta)$	The function $\log\left(\frac{\theta}{1-\theta}\right)$ for $\theta \in (0,1)$
m_*, M_*	Lower and upper boundary on the sample size in stratum $*$
M_{1*}, M_{2*}	Components of the MSE decomposition used in general EBP approaches
$\mathbf{m}$	Arbitrary vector
$\mu, \mu_*, \boldsymbol{\mu}$	Mean, mean of area $*$, vector of small area means
$\widehat{\mu}_*^{**}$	Estimator $**$ for the mean of area $*$
$\text{MSE}(*)$	MSE of $*$
$\widehat{\text{MSE}}(*)$	MSE estimator of $*$
n,n_*	Sample size, sample size in area $*$
N,N_*	Population size, Population size of the area $*$
$\widehat{N}_*$	Estimated population size in area $*$
$N(\mu, \sigma^2)$	Normal distribution with expectation μ and variance σ^2
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with expectation $\boldsymbol{\mu}$ and variance-covariance matrix Σ
ν	Arbitrary variable
O	Order of magnitude in probability
ω	Ratio of the variance components
p_*	Either the selection probability for unit $*$ or the success probability in a Bernoulli trial for unit $*$
P_*	Weight for area $*$ in the power and Longford allocation
π_*	Inclusion probability of observation $*$
$\pi_{*\#}$	Joint inclusion probability of observations $*$ and $\#$
π_L, π_U	Lower and upper bound of box-constraint inclusion probabilities
$\Pr(*)$	Probability of event $*$
ψ_*	Variance of the direct estimator in area $*$
q	Arbitrary constant
r, R	Running index $r = 1 \ldots R$
$\mathbf{R}$	Variance-covariance matrix of the sampling error
$\mathcal{S}$	Set of all possible samples
S, S_*	Sample, sample in area $*$
$S^2_{*,**}$	Variance of the variable $*$ in stratum $**$
σ^2_*	Variance of the variable $*$
SSB	Sum of squares between PSUs
SSW	Sum of squares within PSUs
SSTOT	Total sum of squares
τ, τ_*	Total, total in area $*$
$\widehat{\tau}_*^{**}$	Total estimate from estimator $**$ in area $*$
θ, θ_*	Arbitrary statistic, arbitrary statistic in area $*$
U, U_*	Universe or population, universe or population in area $*$
v_*	Domain-specific effect for domain $*$
$\mathbf{v}$	$q \times 1$ vector of the random effects
$\mathbf{V}$	$n \times n$ variance-covariance matrix on sample length

$\mathrm{Var}(*)$	The variance of $*$
w_*	Design weight for unit $*$
$\mathbf{X}$	$n \times p$ matrix of the covariates including the preceding column of ones on sample length
$\mathbf{x}_*$	Vector of auxiliary information for unit $*$
$\boldsymbol{\xi}$	Vector of the model parameters
$\mathbf{y}$	$n \times 1$ vector of the dependent variable on sample length
y_*	Variable of interest for unit $*$
$\widehat{y}_*$	Predicted value of the variable of interest for unit $*$
$\overline{y}, \overline{y}_*$	Sample mean, sample mean in PSU $*$
$\overline{Y}, \overline{Y}_*$	Population mean, population mean in PSU $*$
$\boldsymbol{z}_*$	Vector of auxiliary information related to the sampling design for unit $*$
z	Poverty threshold
ζ_*	Transformed variable of interest for unit $*$

List of Abbreviations

ACR	Average coverage rate
AIC	Akaike information criterion
ARPR	At-risk-of-poverty rate
ARRMSE	Average relative root mean squared error
ATC	Antithetic clustering
BHF	Battese-Harter-Fuller
BLUE	Best linear unbiased estimator
BLUP	Best linear unbiased predictor
BP	Best predictor
EBLUP	Empirical best linear unbiased predictor
EBP	Empirical best predictor
EU-SILC	EU survey on income and living conditions
FH	Fay-Herriot
GREG	Generalised regression
HB	Hierarchical Bayes
HT	Horvitz-Thompson
ICC	Intraclass correlation coefficient
JLW	Jiang-Lahiri-Wan
LFS	Labour force survey
LGREG	Logistic generalised regression
LR	Lohr-Rao
MARB	Mean absolute relative bias
ML	Maximum-likelihood
MSE	Mean squared error
MultGREG	Multinomial logistic generalised regression
OLS	Ordinary least squares
πps	Probabilities proportional to size
PS	Pfeffermann-Sverchkov
REML	Restricted maximum-likelihood
RRMSE	Relative root mean squared error
PSU	Primary sampling unit
QQ	Quantile-quantile
SIC	Single stage cluster sampling
SRS	Simple random sampling
StrRS	Stratified random sampling
SWEE	Survey-weighted estimating equations
UPS	Unequal probability sampling
WOR	Without replacement
WR	With replacement

Chapter 1

Applying small area estimation in a design-based framework

Surveys are commonly tailored to produce estimates of aggregate statistics with a desired level of precision. This may lead to very small sample sizes for subpopulations of interest, defined geographically or by content, which are not incorporated into the survey design. In accordance with Rao (2003, p. 1), we refer to subpopulations where the sample size is too small to provide direct estimates with adequate precision as small areas or small domains. Despite the small sample sizes, reliable small area estimates are needed for economic and political decision making as pointed out by Jiang and Lahiri (2006b, p. 2). Hence, model-based estimation techniques are used which increase the effective sample size by borrowing strength from other areas to provide accurate information for small areas (cf. Rao, 2003, pp. 1; or Jiang & Lahiri, 2006b, pp. 3).

The paragraph above introduced small area estimation as a field of survey statistics where two conflicting philosophies of statistical inference meet: the design-based and the model-based approach. While the first approach is well suited for the precise estimation of aggregate statistics, the latter approach furnishes reliable small area estimates. In most applications, estimates for both large and small domains based on the same sample are needed. This poses a challenge to the survey planner, as the sampling design has to reflect different and potentially conflicting requirements simultaneously. In order to enable efficient design-based estimates for large domains, the sampling design should incorporate information related to the variables of interest. This may be achieved using stratification or sampling with unequal probabilities. Many model-based small area techniques require an ignorable sampling design such that after conditioning on the covariates the variable of interest does not contain further information about the sample membership. If this condition is not fulfilled, biased model-based estimates may result, as the model which holds for the sample is different from the one valid for the population. Hence, an optimisation of the sampling design without investigating the implications for model-based approaches will not be sufficient. Analogously, disregarding the design altogether and focussing only on the model is prone to failure as well. Instead, a profound knowledge of the interplay between the sample design and statistical modelling is a prerequisite for implementing an effective small area estimation strategy.

In this work, we concentrate on two approaches to address this conflict. Our first approach takes the sampling design as given and can be used after the sample has been collected.

It amounts to incorporate the survey design into the small area model to avoid biases stemming from informative sampling. Thus, once a model is validated for the sample, we know that it holds for the population as well. We derive such a procedure under a lognormal mixed model, which is a popular choice when the support of the dependent variable is limited to positive values. Besides, we propose a three pillar strategy to select the additional variable accounting for the design, based on a graphical examination of the relationship, a comparison of the predictive accuracy of the choices and a check regarding the normality assumptions.

Our second approach to deal with the conflict is based on the notion that the design should allow applying a wide variety of analyses using the sample data. Thus, if the use of model-based estimation strategies can be anticipated before the sample is drawn, this should be reflected in the design. The same applies for the estimation of national statistics using design-based approaches. Therefore, we propose to construct the design such that the sampling mechanism is non-informative but allows for precise design-based estimates at an aggregate level.

The remaining chapters are organised as follows:

In Chapter 2, we review design-based and model-assisted strategies for domain estimation. These procedures often combine the choice of a sampling design and an estimator to produce accurate estimates for large domains, which are already incorporated in the sampling design.

In Chapter 3, we present model-based small area estimation approaches that are applicable when the sample sizes for the subgroups of interest are very small. Our focus in this chapter is on normal as well as lognormal mixed models. Afterwards, various tools for model validation are introduced.

Chapter 4 is devoted to extensions of model-based small area estimation procedures to account for the sampling design. After a review of approaches that are suitable for normal mixed models, we develop an empirical best predictor under an augmented lognormal mixed model. We compare this estimator and competing methods in model-based simulation studies under informative sampling.

In Chapter 5, we investigate the trade-off the survey planner faces when a sampling design should be constructed that provides efficient design-based estimates on aggregates as well as permitting model-based small area estimates. To solve this conflict, we propose to select a simple random sample of clusters which are constructed to maximise heterogeneity within clusters with respect to a correlated auxiliary variable. Simulation studies are used to compare our proposal against other sampling designs in a variety of situations.

Chapter 6 presents selected applications of small domain estimation in a design-based environment. A wide range of specialised small area estimation problems is covered, ranging from business statistics over labour market characteristics to poverty measures.

The insights of this work are summarised and an outlook on future research areas is given in Chapter 7.

Appendix A gives a brief overview regarding different simulation strategies in small area estimation. Appendices B and C contain additional material for Chapters 4 and 5, respectively.

Chapter 2

Design-based and model-assisted domain estimation

In this chapter, design-based strategies used in surveys to produce national estimates as well as figures for subgroups are presented. Our exposition starts with the definition of basic concepts used in design-based estimation in Section 2.1. This paves the way for introducing commonly used estimation techniques such as the Horvitz-Thompson estimator or the generalised regression estimator in Section 2.2. Section 2.3 is devoted to probability sampling designs that are applicable for standard surveys, where the sole aim is to estimate aggregate statistics. Extensions of these procedures to the case where additionally estimates for subgroups are desired are presented in Section 2.4. Finally, a discussion of these strategies is given in Section 2.5.

2.1 Concepts in design-based estimation

2.1.1 Notations and definitions

In the following, we will introduce some terminology and concepts important for design-based estimation. A thorough overview about design-based estimation methods for small domains is given by Lehtonen and Veijanen (2009), and unless otherwise stated, our exposition in the following is based on their work. We shall assume that a fixed and finite population U of size N exists. The population is indexed by k with $k = 1, \ldots, N$ and our variable of interest is $y_k,\ k = 1, \ldots, N$. Typically, in design-based estimation the production of a statistic $\theta = f(y_k)$ for the population U is of utmost importance, where the statistic is a function of the variable of interest in the population. This would be very easy if y_k were known for all $k \in U$. In practice, however, the desired information is known for the sample $S \subset U$ of n elements only. The particular sample S belongs to the set of all possible samples $\mathcal{S}$. This leads to the sample distribution, which is defined by the probability mass function $P(\cdot)$ defined on $\mathcal{S}$ and satisfies $P(S) \in [0, 1]\ \forall S \in \mathcal{S}$ and $\sum_{S \in \mathcal{S}} P(S) = 1$ (cf. Särndal, Swensson, & Wretman, 1992, Section 2.3). A special feature of the design-based framework is that the sample is collected by means of probability sampling, i.e. each unit k in the population has a known and positive probability of being included in the sample (cf. Särndal et al., 1992, p. 32). An important concept in the

design-based framework is the probability that a particular unit is included in the sample, which is also known as the (first-order) inclusion probability π_k. It is given by (cf. Fuller, 2009, Section 1.2):

$$\pi_k = \Pr(k \in S) = \sum_{S \in \mathcal{S}} P(S)\mathbb{I}(k \in S), \tag{2.1.1}$$

where $\mathbb{I}(A)$ is the indicator function, which takes the value 1 if condition A is met and 0 otherwise. Hence, the first-order inclusion probability is simply the sum over the probabilities for all samples in which unit k is included. In a similar vein, the joint inclusion probability that units k and l are both included in the sample is defined as

$$\pi_{kl} = \Pr(k, l \in S) = \sum_{S \in \mathcal{S}} P(S)\mathbb{I}(k \in S)\mathbb{I}(l \in S). \tag{2.1.2}$$

It is easily seen from (2.1.2), that $\pi_{kk} = \pi_k$ and $\pi_{kl} \leq \min(\pi_k, \pi_l)$. The inclusion probabilities are also of crucial importance when constructing design-based estimators, where the design weights defined as the inverse inclusion probabilities $w_k = \pi_k^{-1}$ are frequently used (cf. Lehtonen & Veijanen, 2009, Section 2.1). These basic design weights may be further calibrated to adjust for non-response or to reproduce known aggregate values for some auxiliary information (cf. Särndal, 2007). Those adjustments are the reason why according to Gelman (2007, p. 153), the design weights generally capture much more adjustments due to the complex nature of conducting samples in practice than is reflected by differing selection probabilities. In this work, we shall make the simplifying assumption that non-response is not present, such that w_k is indeed the design weight of unit k.

As pointed out by Lehtonen and Veijanen (2009, p. 221), the randomness in the design-based framework is only due to the selection of the units in the sample. This is in contrast to the model-based framework, where the population is treated as a random realisation from the superpopulation model. Thus, we may define the random variables $I_k = \mathbb{I}(k \in S)$ and $I_{kl} = \mathbb{I}(k, l \in S)$ to represent the sample memberships. Simple calculations yield $\mathrm{E}(I_k) = \pi_k$ and $\mathrm{E}(I_{kl}) = \pi_{kl}$. Hence, estimators are evaluated based on the set of all possible samples, which is also known as the (design-based) randomisation distribution. This point will be very important later on, since in practice, all we have is one particular sample. The expectation of an estimator $\widehat{\theta}$ is given by (Särndal et al., 1992, Section 2.7)

$$\mathrm{E}\big(\widehat{\theta}\big) = \sum_{S \in \mathcal{S}} P(S)\widehat{\theta}(S). \tag{2.1.3}$$

An estimator is design-unbiased, if its design bias is zero, which implies that $\mathrm{E}\big(\widehat{\theta}\big) = \theta$. Furthermore, to assess the precision of design-unbiased estimators, the design variance is used. It is given by

$$\mathrm{Var}(\widehat{\theta}) = \sum_{S \in \mathcal{S}} P(S) \left(\widehat{\theta}(S) - \mathrm{E}\left(\widehat{\theta}\right)\right)^2. \tag{2.1.4}$$

If an estimator with a design bias is used, we should instead use the mean squared error (MSE), which considers the squared bias as well, defined as

$$\mathrm{MSE}(\widehat{\theta}) = \sum_{S \in \mathcal{S}} P(S) \left(\widehat{\theta}(S) - \theta\right)^2 = \mathrm{Var}(\widehat{\theta}) + \left(E\left(\widehat{\theta}\right) - \theta\right)^2. \tag{2.1.5}$$

2.1.2 Planned and unplanned domain structures

In domain estimation, we are interested in estimating certain statistics simultaneously for many subgroups of the population. In the following, we will use the terms domain and area interchangeably. Specifically, we consider D mutually exclusive and exhaustive domains with $U_d \subset U, \quad d = 1, \ldots, D$, where $U_1 \cup \ldots \cup U_d \cup \ldots \cup U_D = U$. A direct consequence of this is that each unit of the population belongs to exactly one domain d. The domain sizes are given by N_d and satisfy $\sum_{d=1}^{D} N_d = N$. Note that the domain sizes might be unknown at the design stage, i.e. the domain membership is known only after the unit has been sampled. The part of the sample which is taken from domain d is $S_d = S \cap U_d$ and the resulting sample size is denoted as n_d. The (possibly random) domain-specific sample sizes n_d fulfil $\sum_{d=1}^{D} n_d = n$.

An important distinction for domain estimation is whether the domains of interest are planned or unplanned. The concept of planned domains requires that the domain membership is incorporated in the design (cf. Lehtonen & Veijanen, 2009, p. 222), i.e. the domains are strata. In this case, the domains can be viewed as separate sub-populations implying that the domain sizes are known. Moreover, the (expected) sample size n_d is fixed. Hence, provided that the n_d are not too small, direct estimation methods which rely only on information from the particular domain can be applied.

If the domains are unplanned, the domain membership is not incorporated in the survey design. Hence, the sample size n_d is a random variable, which may be even zero in some cases. Obviously, with $n_d = 0$ any direct estimation strategy is not feasible any more and indirect estimation methods, using data from other domains, have to be considered. Moreover, even if $n_d > 0$, unplanned domains often lead to an increase of the variance owing to the random sample sizes. As an example, suppose that by means of a simple random sample without replacement n units are drawn from a population comprising N units. Under these circumstances n_d follows the hypergeometric distribution. The expected sample size is given by

$$\mathrm{E}(n_d) = n \cdot \frac{N_d}{N} > 0,$$

but the probability for the event $n_d = 0$ is strictly greater than zero. This should be reflected in the estimation of the variance. For large domain sizes it might still be the case that $\mathrm{E}(n_d) \gg 0$ and direct estimation methods may be suitable.

It may be noted that in many surveys, where estimates are required on different levels of aggregation we have both types of domains at once. As an example, imagine a German survey in which the federal states are strata and estimates are desired for the states as well

as for the districts within the states. In this setting, the states are planned areas whilst the districts are unplanned areas. Hence, direct estimation methods may be accurate for the states whereas indirect techniques are required to provide reliable estimates for the districts.

2.2 Domain estimators

2.2.1 Design-based estimators

Design-based estimators have been used in the field of survey statistics for a long time. Suppose, we want to estimate the domain total defined as (Lehtonen & Veijanen, 2009, p. 221)

$$\tau_d = \sum_{k \in U_d} y_k = \sum_{k \in U} y_{dk}, \tag{2.2.1}$$

where the latter expression is based on the extended domain variable of interest $y_{dk} = \mathbb{I}(k \in U_d) y_k$ which is equal to y_k if unit k belongs to domain d and zero otherwise. Alternatively, we could as well want to estimate the mean, which follows as $\mu_d = \tau_d / N_d$. Perhaps the simplest estimator of the domain total is the Horvitz-Thompson (HT) - estimator given by

$$\widehat{\tau}_d^{HT} = \sum_{k \in S_d} \frac{y_k}{\pi_k} = \sum_{k \in S_d} w_k y_k, \tag{2.2.2}$$

which is design-unbiased for τ_d (cf. Lehtonen & Veijanen, 2009, p. 226). This property can be easily demonstrated by rewriting (2.2.2) using the sample membership indicator introduced in Section 2.1 as

$$\widehat{\tau}_d^{HT} = \sum_{k \in U_d} I_k \frac{y_k}{\pi_k}. \tag{2.2.3}$$

Since both the y_k as well as the π_k are fixed in the design-based framework, the only random variable present in (2.2.3) is the sample membership indicator I_k, whose expectation is given by π_k. Thus, the HT-estimator is design-unbiased. Furthermore, (2.2.2) is a direct estimator, as it does not make use of any information outside the domain for estimating τ_d. Besides it does not use any auxiliary information at the estimation stage, which is why it may be inefficient in the presence of strong predictive covariates. In many situations, however, (2.2.2) may work well, nonetheless. In fact, provided $\pi_k \propto y_k$ holds for all elements, its variance is zero (cf. Särndal et al., 1992, p. 88). This is an aim of the probability proportional to size (πps) schemes discussed in Section 2.3. Alternatively, (2.2.2) can be replaced by a Hajek-type estimator given by

$$\widehat{\tau}_d^{Hajek} = N_d \sum_{k \in S_d} \frac{w_k y_k}{\widehat{N}_d}, \tag{2.2.4}$$

where $\widehat{N}_d = \sum_{k \in S_d} w_k$ is the estimated domain size. It should be noted that (2.2.4) involves a ratio of random variables and is thus no longer unbiased in general. Despite this, the Hajek estimator (2.2.4) is often the better choice as pointed out by Särndal et al. (1992, p. 183), who argue that under variable sample sizes n_d both the numerator and the denominator of the quotient in (2.2.4) tend to vary in the same direction. Moreover, a poor correlation between the dependent variable and the inclusion probabilities is better dealt

with by using the sum of weights in the denominator. In some commonly used designs with planned domain structures and fixed sample sizes, (2.2.4) collapses to (2.2.2), which requires that $\widehat{N}_d = N_d$. The variance of (2.2.2) follows from the fact that the HT-estimator is a linear function of the random variables I_k by (2.2.3) and thus given by (cf. Lehtonen & Veijanen, 2009, p. 226)

$$\mathrm{Var}(\widehat{\tau}_d^{HT}) = \sum_{k \in U_d} \sum_{l \in U_d} (w_k w_l / w_{kl} - 1)\, y_k y_l \tag{2.2.5}$$

which can be unbiasedly estimated via

$$\widehat{\mathrm{Var}}(\widehat{\tau}_d^{HT}) = \sum_{k \in S_d} \sum_{l \in S_d} (w_k w_l - w_{kl})\, y_k y_l. \tag{2.2.6}$$

To estimate the variance of a Hajek-type estimator, we note that (2.2.4) is a non-linear statistic composed as the ratio of two estimated totals. Hence, the method of Taylor linearisation due to Woodruff (1971) can be applied. This strategy is employed by Lehtonen and Veijanen (2009, p. 227) and leads to an approximation as

$$\mathrm{Var}(\widehat{\tau}_d^{Hajek}) \approx \sum_{k \in U_d} \sum_{l \in U_d} (w_k w_l / w_{kl} - 1)\, (y_k - \mu_d)\, (y_l - \mu_d)\,, \tag{2.2.7}$$

which can be estimated by

$$\widehat{\mathrm{Var}}(\widehat{\tau}_d^{Hajek}) = \sum_{k \in S_d} \sum_{l \in S_d} (w_k w_l - w_{kl}) \left(y_k - \widehat{\mu}_d^{Hajek}\right) \left(y_l - \widehat{\mu}_d^{Hajek}\right), \tag{2.2.8}$$

where $\widehat{\mu}_d^{Hajek} = \sum_{k \in S_d} w_k y_k / \sum_{k \in S_d} w_k$ denotes the Hajek-estimator of the mean. Note that these expressions depend on the second-order inclusion probabilities via $w_{kl} = 1/\pi_{kl}$, which may be difficult to obtain in some cases. Therefore alternative variance estimators avoiding their computation are often desired. Further details regarding this topic can be found in Lehtonen and Veijanen (2009, section 3.1) and references therein. Fortunately, for many of the sampling designs discussed in the following, expressions (2.2.5) and (2.2.6) simplify tremendously.

2.2.2 Model-assisted estimators

In the previous section, we introduced design-based procedures, which did not make explicit use of auxiliary information at the estimation stage, albeit in the case of the Hajek-type estimator the domain size N_d could be viewed as auxiliary information. Note that already at the design stage some information may be incorporated, e.g. in the construction of strata or in determining the inclusion probabilities. Nonetheless, in many cases an increase in precision can be achieved by using auxiliary information at the estimation stage as well. Procedures, which use a model to reduce the variance compared to design-based estimators, but whose design-based properties are not dependent upon the validity of the model are called model-assisted (cf. Särndal et al., 1992, Remark 6.4.1). The most prominent example is the generalised regression (GREG) estimator introduced by Cassel, Särndal, and Wretman (1976). Under this approach the total can be estimated via

$$\widehat{\tau}_d^{GREG,I} = \sum_{k \in U_d} \widehat{y}_k + \sum_{k \in S_d} w_k \left(y_k - \widehat{y}_k\right), \tag{2.2.9}$$

where the $\widehat{y}_k = f(\mathbf{x}_k; \boldsymbol{\xi})$ are the predictions obtained from some model. As input factors the values of the auxiliary variable $\mathbf{x}_k$ and a vector of model coefficients $\boldsymbol{\xi}$ are used. Thus, the model can be viewed as a tool to produce the predictions $\widehat{y}_k$, needed to estimate the total.

Alternatively, we may consider a second variant of the GREG estimator which can be more efficient, given by

$$\widehat{\tau}_d^{GREG,II} = \sum_{k \in U_d} \widehat{y}_k + \frac{N_d}{\widehat{N}_d} \sum_{k \in S_d} w_k \left(y_k - \widehat{y}_k\right). \tag{2.2.10}$$

It should be noted that (2.2.10) comprises the Hajek-type estimator (2.2.4) as a special case, when the assisting model is linear and assumes a common mean for all units within a domain.

The second expression for the GREG estimator is often more desirable, since considering a Hajek-type estimator of the total of the residuals can be more stable with varying sample sizes n_d (cf. Lehtonen & Veijanen, 2009, p. 234). This result is supported by the findings of Hidiroglou and Patak (2004), who established that the bias conditional on the realised sample sizes is almost zero for estimators of type (2.2.10). Furthermore, they conducted a simulation study which revealed that even for moderately small sample sizes around 20, the conditional coverage rates of type II estimators under a ratio model is very close to the nominal rate. These conditional properties are very important in practice, since practitioners are interested in interpreting the current sample at hand rather than in quantities averaged over all possible samples. An advantage of the type-I GREG estimator is that the estimated domain totals automatically add up to the national GREG estimate, if a linear fixed effects model is used (cf. Särndal et al., 1992, p. 399). This is a property which does not hold for the type-II GREG in general. Finally, it should be noted that whenever $\widehat{N}_d = N_d$ expression (2.2.10) reduces to (2.2.9).

An important feature of GREG estimators is that depending on the level at which the assisted model is fitted, the resulting estimator can be either direct or indirect. On the one extreme, consider the case where the model is fitted separately in each domain. In this case, the predictions $\widehat{y}_k$ depend only on information from the particular domain and hence, estimators (2.2.9) and (2.2.10) are direct. Moreover, for some assisting models, the weighted residual sum will be zero by construction, when a direct GREG estimator is used (cf. Lehtonen & Veijanen, 2009, p. 229). On the other extreme, the assisting model can be fitted using the whole sample. This yields an indirect GREG estimator, which may compensate for small sample sizes by borrowing strength from other areas. Moreover, the model can be fitted to groups of areas, for which a common model seems reasonable.

Both expressions, (2.2.9) and (2.2.10) reveal that the GREG estimator comprises two parts. On the one hand, we have the total of the model-based predictions for all elements belonging to the domain. On the other hand, we have a HT or Hajek-type weighted sample total of the residuals within the domain. The role of these weighted residuals for an indirect GREG estimator is to correct for a potential bias due to the use of the model-based predictions (cf. Lehtonen & Veijanen, 2009, p. 229). Thus, if the assisting model is strong, the residuals will be small and the indirect GREG estimator will be driven by the model-based predictions. If the model is rather poor, however, the weighted residuals correct for the shortcomings of the predictions. As a consequence, the GREG is not model-based but model-assisted which implies that its design properties such as

the asymptotic design-unbiasedness do not depend on the validity of the assisting model (cf. Särndal et al., 1992, Remark 6.4.1). This robustness towards model misspecification comes at the price of an increase of the variance. Suppose that the assisting model is linear and does not account for domain-specific variation. In this case, whereas the variance of the synthetic part is of order $O(n^{-1})$ it amounts to $O(n_d^{-1})$ for the weighted residuals and thus also for the GREG estimator (cf. e.g. Pfeffermann, 2013, Section 4.1).

The approximate variance for the type I GREG estimator follows from inserting the residuals e_k instead of the y_k in the variance formula for the HT-estimator (2.2.6) and is given by (cf. Hidiroglou & Patak, 2004)

$$\mathrm{Var}(\widehat{\tau}_d^{GREG,I}) = \sum_{k \in U_d} \sum_{l \in U_d} (w_k w_l / w_{kl} - 1)\, e_k e_l, \tag{2.2.11}$$

where $e_k = y_k - \widehat{y}_k$ and an estimate follows as

$$\widehat{\mathrm{Var}}(\widehat{\tau}_d^{GREG,I}) = \sum_{k \in S_d} \sum_{l \in S_d} (w_k w_l - w_{kl})\, e_k e_l. \tag{2.2.12}$$

Alternatively, variance estimators using g-weights may be used, which can be preferable in practice (cf. Lehtonen & Veijanen, 2009, p. 230).

So far the exposition of the two GREG variants has been fairly general. We did neither specify the functional form $f(\cdot)$ nor did we make any statements about the way the unobserved heterogeneity is accounted for. Moreover, a decision has to be made on whether to include weights in the model fitting or not. Lehtonen, Särndal, and Veijanen (2003, 2005) study the effect of model choice in great detail. It turns out that albeit appealing from a theoretical point of view, modelling the unobserved heterogeneity between domains does not necessarily pay off in large increases of precision when applied to realistic data sets (cf. Lehtonen et al., 2003).

Hidiroglou and Patak (2004) show that under a linear assisting model including an intercept with homoscedastic error terms estimated by pseudo maximum-likelihood fitting the model separately in each domain yields lower approximate variances than a fully combined model. This statement holds for both GREG-types given by (2.2.9) and (2.2.10). Nonetheless, this argument may seem at odds with common reasoning in small area estimation. To fit a separate model in each domain implies that the regression coefficients which enter the GREG formulae via the predicted values are computed solely using data from that particular domain. In other words: The resulting estimator is a direct estimator. With small sample sizes, however, direct estimation techniques are considered unreliable due to a high variance. It is not clear, why the estimated regression coefficients should be reliable in this case. Hence, we may aim to borrow strength by using information from other domains as well, leading to an indirect estimator. The stabilisation of the regression parameters and therefore the fitted values may come at the price of a potential bias of the predictions $\widehat{y}_k$, if the common regression parameters deviate from the domain-specific counterparts. Due to the structure of GREG estimators such a misspecification is compensated by the weighted residuals and hence the impact is not as bad as on model-dependent procedures. Despite this, it could be accompanied by an increase of the variance of the residuals. We should indicate that Hidiroglou and Patak (2004) attenuate their conclusions. Specifically, they note that basing the decision between a indirect or a direct type II - GREG estimator solely on the approximate variance is not always

recommended. They attribute this to a potential bias with small sample sizes. Following the discussion above, it could as well be argued that placing too much emphasis on the asymptotic characteristics in small domain estimation is questionable, as the sample sizes do not justify asymptotic arguments easily (cf. Pfeffermann, 2006, p. 69). Instead, fitting a separate model in each domain may lead to highly variable regression parameter estimates, which could be avoided by estimating the model for groups of domains or the sample as a whole. Moreover, if a rich assisting model with a large number of covariates is employed, the sample size might be too small to fit the model to the data at all.

Regarding the choice between design-weighted and unweighted estimation, we may note that incorporating design weights amounts to estimating population quantities (cf. Korn & Graubard, 1999, Section 4.3), whereas unweighted estimates refer to population quantities only if the design is self-weighting. This benefit of weighted estimation is achieved at the expense of an increase of the variance. Moreover, if the model can be validated and the sampling design can be ignored, unweighted estimates are model-unbiased for population quantities. An additional benefit of using weighted estimation of the regression parameters is that they may lead to a simplification of GREG estimators. This observation is based on the development by Särndal et al. (1992, p. 231) for national estimators under a linear assisting model and has been used by Hidiroglou and Patak (2004) to establish the above mentioned property that a direct GREG estimator has a smaller approximate variance than its indirect counterpart. Concretely, if the bias correction term $\sum_{k \in S_d} w_k(y_k - \widehat{y}_k)$ is zero by construction, the GREG reduces to the first part in (2.2.9) or (2.2.10), which equals the sum of the predictions.

Turning the attention towards to the functional form of the assisting model, depending on the structure of the data many different approaches can be considered. The most commonly used form for the estimation of national statistics is a linear assisting model that does not account for heterogeneity unexplained by the covariates such that

$$\widehat{y}_k = \mathbf{x}_k^T \widehat{\boldsymbol{\beta}}. \tag{2.2.13}$$

If a common model is fitted, this yields the following type-I GREG estimator

$$\begin{aligned} \widehat{\tau}_d^{GREG,I} &= \sum_{k \in U_d} \mathbf{x}_k^T \widehat{\boldsymbol{\beta}} + \sum_{k \in S_d} w_k \left(y_k - \mathbf{x}_k^T \widehat{\boldsymbol{\beta}} \right) \\ &= \widehat{\tau}_d^{HT} + \left(\boldsymbol{\tau}_{X,d} - \widehat{\boldsymbol{\tau}}_{X,d}^{HT} \right)^T \widehat{\boldsymbol{\beta}}. \end{aligned} \tag{2.2.14}$$

To employ the GREG in small area estimation problems, it may be desirable to model the unobserved heterogeneity either by dummy variables or by random effects. Unless the sample sizes are large it may be more reasonable to incorporate the area-specific differences by random effects, a topic which will be elaborated in more detail in Chapter 3 of the thesis. Employing a random intercept model yields predictions as

$$\widehat{y}_k = \mathbf{x}_k^T \widehat{\boldsymbol{\beta}}_{LMM} + \widehat{v}_d, \quad k \in U_d, \tag{2.2.15}$$

where $\widehat{v}_d$ is the prediction for the random effect of domain d and $\widehat{\boldsymbol{\beta}}_{LMM}$ is the vector of estimated regression parameters from fitting the model using random effects for the domains. Alternatively, the model may include also random slopes, which is the case considered by Lehtonen et al. (2003). A highly relevant question is under which circumstances using the predictions (2.2.15) may lead to large efficiency gains compared to using the

predictions (2.2.13) from the simple fixed effects model. It turns out that under planned domain structures which additionally satisfy $\widehat{N_d} = N_d$, the only difference is due to the different estimates for the vector of fixed effects $\boldsymbol{\beta}$. In this case, the predicted random effects cancel out, which will be outlined in the following. To see this, note first that whenever $\widehat{N}_d = N_d$, (2.2.9) and (2.2.10) are equivalent. Inserting (2.2.15) into (2.2.9) yields a multilevel generalised regression (MGREG) estimator

$$\begin{aligned}\widehat{\tau}_d^{MGREG} &= \sum_{k\in U_d}(\mathbf{x}_k^T\widehat{\boldsymbol{\beta}}_{LMM} + \widehat{v}_d) + \sum_{k\in S_d} w_k\left(y_k - (\mathbf{x}_k^T\widehat{\boldsymbol{\beta}}_{LMM} + \widehat{v}_d)\right) \\ &= \sum_{k\in U_d}\mathbf{x}_k^T\widehat{\boldsymbol{\beta}}_{LMM} + \sum_{k\in S_d} w_k\left(y_k - \mathbf{x}_k^T\widehat{\boldsymbol{\beta}}_{LMM}\right).\end{aligned} \tag{2.2.16}$$

The second equality is due to the fact that we considered the case $\sum_{k\in S_d} w_k = \widehat{N}_d = N_d$. Expression (2.2.16) is almost the same as (2.2.14) obtained under a linear fixed effects model. The only difference is that for the MGREG estimator the regression vector is estimated under a random intercept model, whereas for the GREG estimator we use a linear fixed effects model. Suppose further that the random intercept model holds, the sample size is the same in all domains and the sampling design is non-informative, e.g. the variables which determine the sample selection are included in the model. It is a well-known fact that under these - restrictive - conditions the pooled OLS estimate of $\boldsymbol{\beta}$ is model-unbiased (cf. Verbeek, 2008, Chapter 10). Thus, by careful planning of the survey design, the differences due to model variations can be very small in a GREG approach. It should be noted that whether the domains are planned or unplanned also directly affects the variance of the MGREG estimator. To do so, we follow the arguments of Särndal et al. (1992, p. 401). Consider the hypothetical census-fit residuals which are given by

$$E_k = y_k - \mathbf{x}_k^T\boldsymbol{\beta}_{LMM} - v_d, \quad \text{for } k \in U_d. \tag{2.2.17}$$

Note that $\boldsymbol{\beta}_{LMM}$ refers to the population vector of regression coefficients and $\widetilde{v}_d$ is the realized random effect that would result if we fitted the random intercept model using data for the whole population. The total is then given by

$$\tau_d = \sum_{k\in U_d} y_k = \boldsymbol{\tau}_{X,d}^T\boldsymbol{\beta}_{LMM} + N_d v_d + \tau_{E,d}. \tag{2.2.18}$$

Now, the random intercept model is fitted to the sample data, yielding an estimated regression vector $\widehat{\boldsymbol{\beta}}_{LMM}$ and a predicted random effect $\widehat{v}_d$. The MGREG estimator follows as

$$\begin{aligned}\widehat{\tau}_d^{MGREG} &= \sum_{k\in U_d}\widehat{y}_k + \sum_{k\in S_d} w_k(y_k - \widehat{y}_k) \\ &= \boldsymbol{\tau}_{X,d}^T\widehat{\boldsymbol{\beta}}_{LMM} + N_d v_d + \widehat{\boldsymbol{\tau}}_{X,d}^T(\boldsymbol{\beta}_{LMM} - \widehat{\boldsymbol{\beta}}_{LMM}) + \widehat{N}_d(\widetilde{v}_d - \widehat{v}_d) + \widehat{\tau}_{E,d}.\end{aligned} \tag{2.2.19}$$

The error can then be expressed as

$$\widehat{\tau}_d^{MGREG} - \tau_d = \underbrace{(\boldsymbol{\tau}_{X,d} - \widehat{\boldsymbol{\tau}}_{X,d})^T(\widehat{\boldsymbol{\beta}}_{LMM} - \boldsymbol{\beta}_{LMM})}_{\approx 0} + (N_d - \widehat{N}_d)(\widehat{v}_d - \widetilde{v}_d) + \widehat{\tau}_{E,d} - \tau_{E,d}. \tag{2.2.20}$$

Under a planned domain structure some sampling designs within areas (e.g. drawing elements with equal probabilities) satisfy $\widehat{N}_d = N_d$ by construction and the second term vanishes. Hence the error is approximately $\widehat{\tau}_{E,d} - \tau_{E,d}$ and standard residual variance estimators can be applied. In general, however, $\widehat{N}_d \neq N_d$ and the second term should not be ignored. Torabi and Rao (2008) derive an expression for the model-based MSE of a GREG estimator under a general two-level model.

A major advantage of a linear assisting model is that the sum of the fitted values can be expressed as a function of the column-wise sums of the covariates, $\boldsymbol{\tau}_{X,d}$. Therefore, we do not need to know $\mathbf{x}_k$ for every unit in the population, instead knowing the sampled values of the auxiliary variables - to fit the model - and the domain totals is sufficient. This allows to compute the GREG in cases where the totals can be obtained from other sources (Särndal et al., 1992, p. 230). Furthermore, under a linear model the GREG can be motivated as a calibration estimator, such that it can be expressed as a linear estimator of the observed values, with weights that exactly reproduce the known totals for the auxiliary variables when applied to the sample (cf. Särndal et al., 1992, p. 234).

Besides a linear model, more sophisticated models can be employed to yield a better fit to the data. The use of model-assisted approaches for categorical data in survey sampling emerged with Lehtonen and Veijanen (1998). We consider the case of a dichotomous dependent variable first, which can be modelled by a logistic regression model. Other options of link functions for GREG-type estimators assisted by generalised linear models are discussed in Myrskylä (2007, Section 4.5). So we now assume that y_k denotes the binary response for the k-th unit. Then a logistic generalised regression (LGREG) estimator for the domain total under an assisting fixed effects model uses predictions

$$\widehat{y}_k = (1 + \exp(-\mathbf{x}_k^T \widehat{\boldsymbol{\beta}}))^{-1} \tag{2.2.21}$$

where $\widehat{\boldsymbol{\beta}}$ refers to vector of regression parameters obtained from regressing the y_k on the $\mathbf{x}_k$ by means of a logistic model. The LGREG estimator is obtained by inserting the expression (2.2.21) into either (2.2.9) or (2.2.10). As in the linear model, unobserved heterogeneity may be modelled using random effects (cf. Lehtonen et al., 2005). To estimate the variance of the LGREG estimator, Lehtonen and Veijanen (1998) propose to use the residual variance. Alternatively, with an application to cluster sampling designs in mind, Kennel and Valliant (2010) consider an estimating equations approach. Moreover, it is important to note that unlike in the linear case, the LGREG estimator needs the auxiliary information for all units in the population. This is due to Jensen's inequality which states that $\mathrm{E}(g(x)) \neq g(\mathrm{E}(x))$ for any non-linear function $g(\cdot)$ and considering that the LGREG estimator uses the inverse logistic link function. Unlike under the linear assisting model, the LGREG estimator does not offer an alternative derivation from a calibration approach unless all variables are categorical and the model parameters are estimated using pseudo maximum-likelihood (cf. Myrskylä, 2007, p. 52). Lehtonen and Veijanen (1998) propose multinomial logistic regression models to extend the LGREG approach for general categorical dependent variables. The model considered is

$$logit(p_{ik}) = \frac{\exp(\mathbf{x}_{ik}^T \boldsymbol{\beta}_i)}{\sum_{i=1}^{I} \exp(\mathbf{x}_{ik}^T \boldsymbol{\beta}_i)}, \quad i = 1, \ldots, I, \tag{2.2.22}$$

where i refers to the categories. In order to avoid identifiability issues, the coefficients $\boldsymbol{\beta}_1$ are all set to zero, where the choice of the category is arbitrary. This leads to the following predictions (cf. Lehtonen & Veijanen, 1998):

$$\widehat{y}_{ik} = \frac{\exp(\mathbf{x}_{ik}^T \widehat{\boldsymbol{\beta}}_i)}{1 + \sum_{r=2}^{I} \exp(\mathbf{x}_{rk}^T \widehat{\boldsymbol{\beta}}_r)}, \quad i = 1, \ldots, I. \tag{2.2.23}$$

Using (2.2.23) and applying the GREG principle leads to the multinomial logistic GREG (MultGREG). The MultGREG may be applied in labour force surveys (LFS) where typically estimates for the number of employed, unemployed and other categories are desired (cf. Särndal, 2007). Alternatively, $I-1$ separate binomial logistic regression models, where elements belonging to category i are coded as 1 and all elements from other categories $r \neq i$ are coded as 0, can be entertained. A major disadvantage of this approach is that the probabilities for the different categories for each element do not necessarily sum to 1. This drawback does not occur with a multinomial logistic model, which from (2.2.23) yields

$$\sum_{i=1}^{I} \widehat{y}_{ik} = \frac{1 + \sum_{r=2}^{I} \exp(\mathbf{x}_{rk}^T \widehat{\boldsymbol{\beta}}_r)}{1 + \sum_{r=2}^{I} \exp(\mathbf{x}_{rk}^T \widehat{\boldsymbol{\beta}}_r)} = 1. \tag{2.2.24}$$

The property (2.2.24) is highly relevant for model-dependent estimators $\widehat{\tau}_i^{Multinomial} = \sum_{k \in U} \widehat{y}_{ik}$ as it ensures that $\sum_{i=1}^{I} \widehat{\tau}_i^{Multinomial} = N$. Thus, the sum of the totals of the categories is coherent with the size of the population, which is typically known in advance. This kind of calibration against the population size holds for direct GREG estimators using a multinomial logistic regression model.

2.3 Sampling designs for national statistics

2.3.1 Simple random sampling

In simple random sampling (SRS), a probability sample is drawn without subdividing the population in any respect, where all elements have the same inclusion probability, $\pi_k = n/N$ for $k = 1, \ldots, N$. We may distinguish between SRS with replacement (SRSWR), where elements can be drawn more than once and simple random sampling without replacement (SRSWOR), where each element can be drawn only once (cf. Lehtonen & Pahkinen, 2004, p. 24). Since the SRSWOR scheme is much more common in statistics, we will denote it as SRS in the following. A notable exception where SRSWR is used, is the basic Monte-Carlo bootstrap (cf. P. Lahiri, 2003 and references therein). The expressions for the estimators discussed in Section 2.2 also simplify tremendously under SRS. It is easily seen that the HT-estimator of the population total τ_y coincides with the Hajek-type estimator and both reduce to $N\overline{y}$, where $\overline{y}$ denotes the simple sample mean. In surveys with design-based estimation strategies, however, SRS is rarely used. This is due to the fact that SRS can be somewhat impractical and, even more important from a design-based viewpoint, it does not make use of any auxiliary information at the design stage, which can be used to increase the precision of design-based estimates. Nonetheless, SRS plays an important role, as many of the more complicated multi-stage designs use SRS at one or more stages. Moreover, SRS serves as an important benchmark to which the relative efficiency of other designs can be compared (Lehtonen & Pahkinen, 2004,

Section 2.3). A prominent example in this regard is the so called design effect, defined as (Lehtonen & Pahkinen, 2004, p. 15Gabler, Häder, & Lynn, 2006)

$$\text{DEFF} = \frac{\text{Var}(\widehat{\tau})_C}{\text{Var}(\widehat{\tau})_{SRS}}, \tag{2.3.1}$$

where $\text{Var}(\widehat{\tau})_C$ denotes the variance of a total under the design of interest. As pointed out by Lehtonen and Pahkinen (2004, p. 16), a design effect greater than one implies a less efficient sampling design than SRS. Furthermore, it should be noted that for design-based small domain estimation SRS is not very appealing, as it implies unplanned domains, as (at least) the domain-specific sample size is not known in advance. This may be problematic and can even lead to non-sampled domains, for which design-based estimates cannot be computed. A procedure, which is sometimes called a SRS design in the model-based small domain estimation literature consists of allocating a given sample size to the domains and performs SRS to select units within domains. It should be noted, however, that this design is in fact a stratified random sampling with domains as strata. Hence, the domains are planned under this procedure (cf. Lehtonen & Veijanen, 2009). These designs will be discussed in the next section.

2.3.2 Stratified random sampling

To elaborate on stratified random sampling (StrRS), we consider a decomposition of the universe U into L pairwise disjoint non-empty strata, i.e. (cf. Särndal et al., 1992, Section 3.7):

$$U_h \cap U_{h'} = \emptyset \quad \text{for all } h = 1, \ldots, L \quad \text{with } \cup_{h=1}^{L} U_h = U.$$

This process yields L strata, which are also known as primary sampling units (PSUs) and contain N_h units of the population such that

$$\sum_{h=1}^{L} N_h = N.$$

In the simplest case, a SRS is conducted within each stratum h, where n_h units are drawn from the hth stratum, fulfilling $\sum_{h=1}^{L} n_h = n$. In contrast to the SRS scheme described previously, StrRS ensures that units from all strata are present in the sample. If the sample is drawn via SRS within the strata, the inclusion probabilities satisfy

$$\pi_{k \in U_h} = \frac{n_h}{N_h}. \tag{2.3.2}$$

It can be seen from this expression that all units from one stratum share the same inclusion probability and hence the same survey weight. In general, however, units from different strata will not have the same survey weight attached to them. Furthermore, the estimates

of a national total obtained by HT and the Hajek estimator coincide if SRS is applied within a StrRS scheme. An important issue in StrRS is the definition of the strata. Since StrRS is often used to reduce the variance of design-based estimates, the variables used in determining the strata should be related to the variable(s) of interest. As an illustration, suppose that the main aim of the survey is to estimate the total $\tau_y = \sum_{k \in U} y_k$ and a HT estimator is used. Assuming SRS within the strata, the variance of the total emerges as

$$\mathrm{Var}(\widehat{\tau}^{HT}_{StrRS}) = \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h} \frac{N_h - n_h}{N_h - 1}, \tag{2.3.3}$$

where S_h^2 is the population variance of the dependent variable in stratum h (cf. Lehtonen & Pahkinen, 2004, p. 63). Hence, if it were possible to construct the strata such that $S_h^2 = 0$ for all h the variance of the total would be zero. This thought experiment highlights that StrRS will lead to a variance reduction, provided that the within-stratum values of the dependent variable are similar. In practice however, it is not possible to construct strata based on a variable for which information is collected by means of this survey (cf. Särndal et al., 1992, Section 3.7.4). But if the sampling frame contains information about a proxy of the variable of interest, maybe obtained from a register, this information could be used. One might further argue that it would be optimal to have as many strata as units in the population, as this implies stratum sizes $N_h = 1$ for all $h = 1, \ldots, L$ and hence the variance within strata is zero. But by the requirement that we sample elements from all strata, the only possible sample would be a census of the population, whose sampling variance is trivially zero in the WOR case. Note that a census is not completely free of errors, due to non-sampling errors (see Groves & Lyberg, 2010 for a review of the concept of total survey error, which accounts for a variety of errors). To achieve a variance reduction in business surveys for example, the strata are frequently obtained as a cross-classification of the company size, the industry in which it operates and some regional information (cf. Hidiroglou & Lavallee, 2009). This stratification will work well, provided that these variables explain a substantial amount of variation in the dependent variables. In many cases, the choice of stratification is not entirely up to the survey planner, as external regulations may prescribe a certain construction of strata. If we consider the stratification as fixed, the remaining question is how to allocate the sample size to the strata.

The proportional allocation aims to yield the same sampling fraction $f_h \doteq n_h/N_h$ in all strata. The formula is thus (cf. Särndal et al., 1992, p. 107):

$$n_{h,PROP} = n\frac{N_h}{N} \quad , h = 1, \ldots, L. \tag{2.3.4}$$

Obviously, expression (2.3.4) may lead to non-integer values for the sample sizes, which need to be adjusted by rounding. As a consequence, the actual sampling fractions may vary slightly from stratum to stratum. Additionally, the rounded values need to satisfy the constraint that $\sum_h n_h = n$. The proportional allocation guarantees (approximately) ignorable design weights, since all units in the population share the same probability of being included in the sample. Furthermore, (2.3.4) avoids overallocation and is thus always feasible for $n < N$. It has been pointed out by Särndal et al. (1992, p. 108 f.) that except in the rather theoretical case that all stratum means are (almost) identical, given

a total sample size of n the proportional allocation yields smaller variances compared to SRS. In fact, if most of the variation in the dependent variable is between strata, the variance reduction may be substantial. A drawback of the proportional allocation is that it may lead to small stratum-specific sample sizes for smaller strata, which can be an issue if design-based estimates for the strata are desired.

Another commonly used simple allocation procedure for stratified random sampling is the so called equal allocation, where the total sample size is evenly distributed among the strata, i.e.:

$$n_{h,EQ} = \frac{n}{L}, h = 1, \ldots, L. \tag{2.3.5}$$

If the ratio n/L is not an integer, the sample sizes have to be rounded. Furthermore, a requirement is that $n/L \leq N_h$. In cases where this does not hold for at least one stratum, the following modification may be considered instead (cf. Choudhry, Rao, & Hidiroglou, 2012):

$$n_{h,EQ} = \begin{cases} N_h, & h \in L_1 \\ \dfrac{n - \sum_{k \in L_1} N_k}{L - m}, & h \notin L_1 \end{cases}, \tag{2.3.6}$$

with L_1 as the set of stratum indices, where the desired sample size due to equal allocation exceeds the stratum size. The remaining sample size, $n - \sum_{k \in L_1} N_k$ is then divided evenly among the remaining $L - m$ strata, where m is the number of strata where the stratum size is larger than n/L. Note that if in some stratum h' the condition $N_{h'} \geq n/L$ holds with equality or N'_h is only slightly larger than n/L, allocation (2.3.6) might be not feasible as well. In this case, we would set $n_{h'} = N_{h'}$ and distribute the remaining sample size evenly among the strata which do not belong to the set $L_1 \cup h'$. Obviously, this can also be done if there is more than one stratum violating the modified condition. An advantage of the equal allocation is that very small sample sizes in the strata are avoided and hence design-based estimation strategies on the stratum level may be pursued. It will also yield national estimates with high precision, provided the stratum-specific variances do not vary much. If on the other hand, the variation of the variable of interest is highly dispersed between different strata, the equal allocation will be inefficient for the estimation of national quantities.

Traditionally, the efficient estimation at the national level has been the most important target of most surveys. Thus, a straightforward approach is to allocate the sample size to the strata as to minimise the variance of the national mean. It can be shown that minimising the variance of the national mean or total in stratified random sampling leads to the optimal allocation proposed by Neyman (1934) and Tschuprow (1923). The optimisation problem is as follows:

$$\begin{aligned} \arg\min_{n_h} \quad & \mathrm{Var}(\widehat{\tau}_{StrRS}) = \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h} \frac{N_h - n_h}{N_h - 1} \\ \text{s.t.} \quad & \sum_{h=1}^{L} n_h = n. \end{aligned} \tag{2.3.7}$$

Using the Lagrangian, the first order conditions for a minimum are given by

$$\frac{\partial}{\partial n_h} = -\frac{N_h}{N_h - 1}\frac{N_h^2 S_h^2}{n_h^2} + \lambda \stackrel{!}{=} 0 \tag{2.3.8}$$
$$\frac{\partial}{\partial \lambda} = \sum_{h=1}^{L} n_h - n \stackrel{!}{=} 0,$$

where λ denotes the Lagrange multiplier. Solving (2.3.8) for n_h yields

$$n_{h,OPT} = n\frac{N_h S_h \sqrt{\frac{N_h}{N_h - 1}}}{\sum_k N_k S_k \sqrt{\frac{N_k}{N_k - 1}}}. \tag{2.3.9}$$

Since the WOR correction terms $\sqrt{\frac{N_h}{N_h - 1}} \approxeq 1$ for large N_h, they are often omitted, when calculating the optimal allocation. The problem in practice, of course, is that the standard deviation of the variable of interest is not known at the time when the sample allocation is computed (Särndal et al., 1992, p. 106). Hence, proxies have to be used, e.g. values from previous years. Obviously, the quality of the resulting sample sizes depend on the accuracy of the proxy information used. If (2.3.9) is calculated using past values of S_h and meanwhile the variable of interest has been subject to a large shock, the resulting allocation may be far from optimal. An example where the use of the proxy leads to disastrous results will be given in the applications in Chapter 6. Moreover, allocation (2.3.9) is optimal for a specific national mean or total. This approach is generally fine if the survey is conducted with one estimate of utmost importance in mind. But in most surveys there are many different variables which should be estimated with sufficiently high precision. In this case, the allocation (2.3.9) may be suitable for highly correlated variables. It does not, however, generalise easily to multivariate estimation problems where this assumption is not plausible. Besides, in the special case of equal variances within the strata the optimal allocation coincides with the proportional allocation.

2.3.3 Cluster sampling

Stratified random sampling is not the only design that can be used, when the universe can be decomposed into L pairwise disjoint, non-empty PSUs. Another frequently used approach is not to sample some elements in all PSUs, but take a census of a few sampled PSUs (cf. Lohr, 1999, Chapter 5). This procedure is called single stage cluster sampling (SIC). Note that the PSUs can be selected by means of a simple random sample, which we will assume in the following. On the one hand, SIC can be a very practical approach, since it does not require a full list of all units in the population and it may save costs, when the PSUs correspond to geographical entities (see Lohr, 1999, Chapter 5 for a further discussion). On the other hand, SIC will not be very efficient for design-based estimation if the units within clusters are more homogeneous than units from different clusters. To assess the efficiency of SIC, we may compute the design effect given by (2.3.1), which in the case of equally sized PSUs is closely related to the ratio of the variation within PSUs to the total variation (Lohr, 1999, Section 5.2). When cluster sampling is applied in surveys, it is usually under a more complex scheme than SIC. To compensate for the likely efficiency

loss when units within PSUs are very similar, two-stage cluster sampling (TSC) can be performed, where instead of a census of the elements within a PSU, a sample of secondary sampling units is taken. This can be more efficient than SIC if the sampling fraction on the first stage is increased but is usually more expensive. In general, the formulae for variances of design-based methods are more complicated with TSC schemes (cf. Lohr, 1999, Section 5.3). It should be noted that this idea can be generalised to multi-stage designs as well. In practice, multi-stage designs are frequently applied, where the first few stages consist of strata and on the last stage clusters are selected. An advantage of these approaches is that the stratified sampling on the higher levels allows for controlling aggregate margins, whilst the cluster sampling on the last stage avoids too high costs.

2.3.4 Unequal probability designs

The designs discussed so far have in common that, at least on the last stage, the elements within a sampling unit share the same probability of being included in the sample. Dispensing with this assumption, we arrive at a bunch of sampling designs that are referred to as unequal probability sampling designs. While there are many different ideas in unequal probability sampling, a commonly used approach is that the inclusion probability π_k should be proportional to a size variable known at the design stage, z_k (Lehtonen & Pahkinen, 2004, p. 49). Designs fulfilling this property are called probability proportional to size (πps) designs. When the desired sample size is set to n, the inclusion probabilities in πps - designs ideally satisfy

$$\pi_{k,\pi ps} = n\frac{z_k}{\sum_{l=1}^{N} z_l}. \tag{2.3.10}$$

In practice, the π_k may differ from expression (2.1.1) owing to the fact that inclusion probabilities have to be greater than zero and may not exceed one (Särndal et al., 1992, p. 89). The first of these requirements excludes size variables that may take negative values or zero. To overcome this problem a transformed size variable may be used, ensuring this condition is met. If for a certain unit k, (2.3.10) yields a value greater than one, π_k is set to one, i.e. unit k is sampled with certainty. To compute the remaining $N-1$ inclusion probabilities, we take their value of the size variable relative to the sum over values of the size variable, excluding unit k and multiply this quantity with $n-1$, since one element is already included in the sample. To draw a sample with the inclusion probabilities given by (2.3.10) several algorithms have been proposed. A comprehensive account on this issue is given by Tillé (2006). In this work, we will only highlight some aspects of them. It is easily seen that the sum over the inclusion probabilities $\pi_{k,\pi ps}$ in the universe equals n, but this does not necessarily imply that also the realised sample is of size n. Instead, if Poisson sampling is applied under probabilities satisfying (2.3.10), the sample size is random with an expected value that is equal to n (Särndal et al., 1992, Section 3.5).

While πps sampling may lead to valuable efficiency gains if applied with a suitable estimation strategy and a strong size variable, its benefits when applied for small domain applications are less clear to some extent. Part of this issue is due to the fact that in small domain problems the sample size can be very small, such that model-based estimation strategies are required. As will be discussed in subsequent chapters, many model-based procedures do not possess desirable design-based properties. Hence, their applicability

hinges upon the degree to which a model can be validated for the sample and applied to the non-sampled part of the population. Albeit incorporating the design weights into the model can be an option as will be discussed in Chapter 4, this requires knowledge about the weights for the non-sampled units as well under non-linear models. In many cases, the model analyst will not have access to this information. Hence, he may be required to fit a model without weights, which can be problematic as πps sampling tends to lead to highly variable design weights, provided the size variable exhibits large variation. Note that if the size variable does not vary considerably, also the potential efficiency gains using a design-based estimator are limited, as the inclusion probabilities given by (2.3.10) will be rather similar, approaching a SRS scheme in the limit. Münnich and Burgard (2012) demonstrate that a large variation of the survey weights may negatively influence the performance of purely model-based small area approaches in realistic circumstances. Thus, in the presence of possible model misspecification, we may want to constrain the variation of the design weights. Therefore, Burgard, Münnich, and Zimmermann (2014), took the idea from Gabler, Ganninger, and Münnich (2012) to the context of unequal probability sampling and applied box-constraints to the inclusion probabilities π_k. A generalisation of their procedures yields box-constraint inclusion probabilities π_k^* determined as the solution

$$\begin{aligned} \min_{\omega_k} \quad & \sum_{k=1}^{N} G(\pi_k^*, \pi_k) \\ \text{s.t.} \quad & \sum_{k=1}^{N} \pi_k^* = n, \\ & 0 < \pi_L \leq \pi_k^* \leq \pi_U \leq 1, \qquad k = 1, ..., N. \end{aligned} \tag{2.3.11}$$

In (2.3.11) π_L and π_U denote the lower and the upper bound for the new box-constraint inclusion probabilities and $G(\cdot)$ is a distance function. Looking at the above programming problem, we note that we try to change the optimal inclusion probabilities π_k emerging from (2.3.10) as little as possible, while forcing them to the range between π_L and π_U. Note that there are many possibilities regarding the choice of $G(\cdot)$, e.g. the distance functions employed in the context of calibration estimation. A reasonable choice is to consider convex functions, ensuring that the problem (2.3.11) can be solved by convex optimisation techniques, which are readily available. In a simulation study, Burgard et al. (2014) demonstrate that using the box-constraint inclusion probabilities stemming from (2.3.11) may yield better results with model-based estimators compared to choosing the π_k via (2.3.10).

2.4 Sampling designs for domain estimation

One major problem of the traditional StrRS designs described in Section 2.3 is that they are either specifically tailored for estimation at the domain level (e.g. equal allocation) or at the national level (e.g. proportional or optimal allocation). In a real world problem, however, the survey designer might not be willing to put up with losses in estimation quality at either level. Hence, there is a need for designs which enable good estimates at both the domain and national level. A natural approach to achieve good estimates is to combine a design suitable for domain level estimation with another one designed for national statistics. This approach has been taken by Costa, Satorra, and Ventura (2004)

who propose a convex combination of the proportional and equal allocation. Assuming a StrRS design where the domains are strata they propose the following sample sizes:

$$n_{h,Costa} = cn_{h,Prop} + (1-c)n_{h,Equal}\,, \quad 0 \leq c \leq 1, \quad h = 1, \ldots, L \tag{2.4.1}$$

where c denotes the share allocated to the proportional allocation and the assumption is made that the strata coincide with the domains of interest. It should be noted that (2.4.1) offers a compromise between desired quality at the national and the domain level, where c reflects the relative importance which is given to the different targets. In a similar fashion, Chiodini, Martelli, Manzi, and Verrecchia (2010) propose a convex combination of the equal and optimal allocation.

Whereas convex combinations of allocations for StrRS, such as the Costa allocation (2.4.1) provide simple solutions for practitioners, they lack a theoretical justification beyond their plausibility. This shortcoming may be dealt with by specifying an objective function that is explicitly minimised. One such approach towards balancing the needs for providing good estimates at different levels of aggregation is given by the so called power allocations. This idea dates back to Bankier (1988), who used power allocations to produce reliable direct estimates at national level and for subgroups determined by strata. They solve the following programming problem:

$$\begin{aligned} \arg\min_{n_h} \quad & \sum_{h=1}^{L} (P_h^q \cdot \mathrm{CV}(\hat{\tau}_{y,h}))^2 \\ \text{s.t.} \quad & \sum\nolimits_{h=1}^{L} n_h = n, \end{aligned} \tag{2.4.2}$$

where P_h reflects the importance of stratum h, $\mathrm{CV}(\hat{\tau}_{y,h})$ denotes the coefficient of variation of the stratum total and q is a constant with $0 \leq q \leq 1$. (2.4.2) can be solved using the Lagrangian multiplier approach. Setting the partial derivatives equal to zero and solving for the stratum-specific sample size n_h yields (Bankier, 1988):

$$n_h = n \frac{S_h P_h^q / \mu_{y,h}}{\sum_{h'=1}^{L} S_{h'} M_{h'}^q / \mu_{y,h'}}, \tag{2.4.3}$$

with $\mu_{y,h}$ and S_h as the mean and standard deviation of the target variable within the hth stratum, respectively. Note that for $P_h = \tau_{y,h}$ and $q = 1$ (2.4.3) reduces to the optimal allocation (2.3.9). This is an interesting property, since the optimal allocation emerges from minimising the variance or coefficient of variation of the national total, whereas the power allocation minimises a weighted sum of the domain-specific coefficients of variation. Alternatively, if $q = 0$, the ratio of the stratum-specific variances of the dependent variable to the stratum-specific means is roughly constant, and the finite population correction is small, similar coefficients of variation will result for the different strata (cf. Bankier, 1988). Thus, the choice of q implicitly defines the relative importance which is given towards estimation of national statistics relative to domain statistics. While $q = 0$ will yield good estimates on the domain level, it may be inefficient for national statistics. As q approaches 1, we get closer to the optimum of the national variance, but the quality of estimates in domain with little importance for the national statistics may deteriorate. Furthermore, as demonstrated by Bankier (1988) choosing $q = 0$ may lead to highly different response burdens in the domains, which could be considered unacceptable in some cases.

The discussion above points out that the parameter q serves two purposes in the power allocation at once. On the one hand it allows to control the degree of smoothing over the

coefficients of variation in domains whereas on the other hand it also reflects the relative importance attributed towards the estimation of national statistics. As discussed by Longford (2006), this dual role of the resulting size measure P_h^q, which he calls "inferential priorities", may be inconvenient. He addresses this issue by introducing an additional parameter G reflecting the relative importance of the national estimates. It may be noted that his exposition is based on the assumption that the domains coincide with the strata. Thus, he proposes a strategy which minimises a weighted sum of the variances of the estimated strata means and the variance of the estimated national mean:

$$\begin{aligned} \arg\min_{n_h} \quad & \sum_{h=1}^{L} P_h \mathrm{Var}(\widehat{\mu}_h) + G P_+ \mathrm{Var}(\widehat{\mu}) \\ \text{s.t.} \quad & n = \sum_{h=1}^{L} n_h. \end{aligned} \tag{2.4.4}$$

In the second part of (2.4.4), G denotes the relative priority given to the estimation of the national mean and P_+ refers to the sum of the stratum-specific priorities, i.e. $P_+ = \sum_{h=1}^{L} P_h$. The programming problem (2.4.4) may be solved by means of the Lagrange-multiplier technique. In the case of general P_h and general sampling designs there exists no closed form solution for the optimal sample sizes. Longford proposes weights of the form $P_h = N_h^q$ with $0 \leq q \leq 2$. In the case of simple random sampling within the strata the choice $q = 2$ leads to the Neyman allocation (2.3.9). Longford (2006) uses the framework of the programming approach (2.4.4) to develop an optimal strategy which considers the sampling design and the estimator simultaneously.

Another way to tackle the problem of stratum and national level estimation at the same time is by determining the minimal sample sizes which satisfy certain quality requirements on the estimates. This approach is taken by Choudhry et al. (2012), who minimise the total sample size subject to constraints of the coefficients of variation at stratum and national level.

$$\begin{aligned} \arg\min_{f_h} \quad & \sum_{h=1}^{L} f_h N_h \\ \text{s.t.} \quad & \mathrm{CV}(\widehat{\mu}_{y,h}) \leq \mathrm{CV}_{0h}, \quad h = 1, \ldots, L \\ & \mathrm{CV}(\widehat{\mu}_y) \leq \mathrm{CV}_0 \\ & 0 < f_h \leq 1, , \quad h = 1, \ldots, L, \end{aligned} \tag{2.4.5}$$

where CV_{0h} denotes the upper boundary on the coefficient of variation for stratum h, while CV_0 refers to the upper boundary allowed for the national mean. To get a convex separable objective function, Choudhry et al. (2012) rewrite the loss function in (2.4.5) as $\sum_{h=1}^{L} k_h^{-1} N_h$, where k_h is the inverse sampling fraction in stratum h and minimise over the choice of the k_h. Furthermore, they re-express the constraints in terms of the squared coefficient of variation, since these expressions are linear in k_h. Choudhry et al. (2012) further consider extensions for composite estimators without covariates and the case where domains cut across strata. The programming problem (2.4.5) can be solved by means of the same methods as problem (2.4.4).

A major difference between the approaches of Longford (2006) and Choudhry et al. (2012) is that the former takes the total sample size as given and minimises a weighted combination of the variances, whereas in the latter approach certain quality thresholds have to be met and the goal is to minimise the sample size. Moreover, the approach due to

Choudhry et al. (2012) is much simpler to implement as the survey planner only needs to specify tolerances on the stratum and national coefficients of variation. To apply the approach due to Longford (2006), however, the inferential priorities as well as the relative importance of the national mean have to be specified.

A wholly different strategy has been proposed by Gabler et al. (2012) in the context of the register-assisted German census 2011. The authors note that the optimal allocation (2.3.9) is efficient for the estimation of the total number of inhabitants in Germany. But with the census also estimates of the population size for geographically defined subgroups, such as federal states, districts and municipalities were desired. An allocation optimal for the national level may yield sample sizes that are too small for direct estimation methods at lower levels of aggregation. In the application of the German census, however, the relative standard error was required to be less than 0.5 % for municipalities with at least 10000 inhabitants by law. Additionally, the law required that no more than 7.9 million people should be included in the sample. Moreover, a highly different treatment of people in terms of the response burden was not desired, either. Hence, the programming problem (2.3.7) had to be modified, to incorporate lower and upper boundaries via box-constraints on the sample size in the different strata. Proceeding as Gabler et al. (2012), the programming objective can be written as

$$\begin{aligned} \arg\min_{n_h} \quad & \mathrm{Var}(\widehat{\tau}_{StrRS}) = \sum_{h=1}^{L} \frac{N_h^2 S_h^2}{n_h} \frac{N_h - n_h}{N_h - 1} \\ \text{s.t.} \quad & 0 < m_h \leq n_h \leq M_h, \quad h = 1, \ldots, L \\ & \sum_{h=1}^{L} n_h \leq n, \end{aligned} \tag{2.4.6}$$

where m_h and M_h denote the lower and upper boundary on the sampling size in stratum h. Efficient numerical algorithms to solve problems such as (2.4.6) are given by Münnich, Sachs, and Wagner (2012). The solution to (2.4.6) is given by

$$n_h^{GGM} = \begin{cases} m_h & \text{for } h \in L_1 \\ \left(n - \sum\limits_{h \in L_1} m_h - \sum\limits_{h \in L_2} M_h\right) \dfrac{N_h S_h \sqrt{\frac{N_h}{N_h - 1}}}{\sum\limits_{k \in L_{12}} N_h S_h \sqrt{\frac{N_k}{N_k - 1}}} & \text{for } h \in L_{12} \\ M_h & \text{for } h \in L_2 \end{cases} \tag{2.4.7}$$

It is evident from (2.4.7) that the solution splits the strata into three groups. For the first group of strata with stratum indices $h \in L_1$, the sample size according to the unconstrained optimal allocation (2.3.9) would be smaller than the lower boundary and hence the sample size is set to this boundary. For a second group of strata with indices $h \in L_2$, the allocated sample size due to (2.3.9) exceeds the upper boundary M_h and hence n_h is set to that boundary. The sample size that is left for the remaining strata, $n - \sum\limits_{h \in L_1} m_h - \sum\limits_{h \in L_2} M_h$ is determined according to (2.3.9). It is interesting to note that in the German census 2011, the strata were constructed within each sampling point, which corresponds to a geographical entity such as a city district in large cities or a municipality with more than 10000 inhabitants (see Münnich, Gabler, Ganninger, Burgard, & Kolb, 2012, Section 2.1 for details). Since domain estimates in the census were required on the level of the sampling points, this structure yields planned domains. The estimator used

was a direct GREG estimator. As the planned domain structure guaranteed sufficiently large sample sizes at the domain-level and the correlation between the covariate and the variable of interest was given by the German Federal Statistical Office to be 0.993 in all sampling points, this strategy yields reliable domain estimates. Münnich, Gabler, et al. (2012) conducted simulation studies based on data from the census test, which indicated that under the assumption of a correlation of 0.993, the quality requirement regarding the relative standard error for municipalities with 10000 inhabitants and more would be met. Owing to the fact that the sampling was conducted independently in each sampling point, the sampling points can be viewed as strata for estimates on higher level aggregates. Due to the peculiarities of a stratified random sample, this ensures that the relative standard error does not exceed 0.5 per cent on aggregations as well (cf. Münnich, Gabler, et al., 2012, p. 25). Furthermore, the box-constrained optimal allocation (2.4.7) also controls the variation of the base-line design weights, since these are given as $w_h = N_h/n_h$ in StrRS. This issue can be important if statistical models are used. Thus, implicitly, the specification of the box-constraints allows to balance the needs between efficient design-based estimates for aggregates and the ability to use model-based estimators for small domains.

2.5 Summary and discussion

Several design-based and model-assisted strategies for small domain estimation have been discussed in this chapter. It should be emphasised at this point that even purely design-based estimation strategies such as HT- or Hajek-type procedures can be used to provide reliable domain estimates under certain conditions (cf. Marker, 2001). In general, this requires a careful survey setup, such that the domains for which estimates are required are known in advance. This ensures that a prespecified, sufficiently large sample size is achieved in each domain. Furthermore, since these design-based estimators do not make use of auxiliary information at the estimation stage, relevant knowledge should be included at the design stage. Traditionally, this has been achieved by optimising the design with respect to the estimation of national statistics, discussed in Section 2.3. Even the explicit consideration of domain-estimates can be easily achieved by including them in the loss function to be minimised, as outlined in Section 2.4. A very interesting proposal in this regard is due to Falorsi and Righi (2008), who consider a two-step procedure ensuring a balanced design. Their approach amounts to derive optimal inclusion probabilities in the first step, such that the anticipated variance respects certain constraints. The second step then consists of manipulating these inclusion probabilities such that the expected sample sizes are calibrated against desired sample sizes in the domains. This design explicitly aims at controlling the sample sizes of domains in cases where the domains are cutting across sampling strata. Their design is somehow related to the one proposed by Choudhry et al. (2012), because Falorsi and Righi (2008) also minimise the total sample size under certain quality constraints in the first step. In the second step, however, they sacrifice the optimality in order to remove any variation of the domain-specific sample sizes.

Moreover, the use of assisting models helps to reduce the variance as they consider sample information other than the variable of interest and use it to smooth out peculiarities of the sample at hand. Additionally, they offer an option to borrow strength in the case of too small sample sizes, by fitting a common model on a group of domains, rather than one domain only. Model-assisted estimators are a very flexible tool in the sense that they allow for a wide range of modelling options. In the practice of survey statistics they are very

popular as they can provide efficiency gains if the assisting model has some predictive power, but in the case of a misspecified model, their design-based properties are not affected. As discussed by Pfeffermann (2013), this protection against model failure leads to asymptotically design-unbiased estimates at the expense of an increase of the variance. Thus, with small sample sizes, the variance of a GREG-type estimator may still be too high to allow for reliable domain estimates. Besides, in the case of non-sampled domains it is not quite clear, how the GREG estimator should be defined. A sample size of 0 implies that there are no observations y_k from that domain and hence applying formulae (2.2.9) or (2.2.10) reduce to the sum of the predictions in this domain. Using an estimator based solely on the predictions, however, is no longer model-assisted but model-based, as the predictions are derived under some model. Hence such an estimator would no longer be asymptotically design-unbiased and we could argue that since the GREG estimator belongs to the group of model-assisted estimators it should not be defined under these circumstances.

Nonetheless, the variance reduction possible by using a design-based or model-assisted estimator may be restricted in practice. As an example, efficient designs may require a very different treatment of the population units in terms of the response burden, beyond of what is deemed acceptable by the society. Besides, approaches to achieve optimal designs for a GREG estimator often require detailed knowledge about the details of the assisting model, such as the variance structure, which may or may not be present (cf. Särndal et al., 1992, Chapter 12 for details). In the worst case, when the assumed knowledge about the model details turns out to be wrong, instead of the desired variance reduction even a increase of the variance compared to simpler approaches may result. However, if the assumed knowledge is accurate, very precise results using an optimal strategy of sample design and estimation can be achieved as the example of the German census 2011 indicates (cf. Münnich, Gabler, et al., 2012).

Altogether, this suggests to incorporate the potential domains of interest already in the sample designs, such that the domain-specific sample size is not to small and fixed. Since the budget of a survey is usually limited and there is a possibly large number of domains of interest, this will not necessarily be feasible. Fuller (1999, p. 344) further noted:

> The client will always require more than is specified at the design stage. For example, the client will explain that they require estimates only at the regional or national level and then, when data are available, ask for county estimates.

This statement illustrates that despite careful planning estimates are often desired for domains which were not planned at the design stage. Since the variances of both design-based and model-assisted strategies are too large to provide estimates with sufficient precision for these unplanned domains, alternatives are needed. The model-based procedures introduced in the next chapter are often better suited for this kind of small domains.

Hence, in modern surveys a variety of different estimation strategies is used simultaneously. For national statistics estimates are produced using design-based or model-assisted (and by definition direct) methods. They are accompanied by estimates stemming from direct or indirect model-assisted method for planned domains with moderately large sample sizes. Finally, in domains with smaller sample sizes and unplanned structures, model-based estimators are employed. This prevailing mix of different strategies has also been criticised as a "design/model compromise" by Little (2012), who raises further issues that

have to be dealt with. One topic which is of utmost importance for statistical agencies is the coherence of estimates. Users require that small domain estimates when aggregated conform with figures for higher aggregation levels. This property is typically not enjoyed by model-based procedures without further adjustments. Another requirement is with respect to the survey design, which besides allowing efficient estimates for design-based and model-assisted procedures has to permit model-building to produce model-based estimates for small domains as well. This introduces additional complexity at the design stage, since design optimisation may lead to the issue of informative sampling (cf. Pfeffermann & Sverchkov, 2009), if not all the variables determining the design can be included in the model. Different options to account for an informative sampling mechanism will be outlined in Chapter 4.

Chapter 3

Model-based small area estimation strategies

As already discussed, for unplanned domains or areas with small sample sizes, direct design-based methods do not provide reliable domain estimates. Hence, there is a need to borrow strength by using information from other domains as well through implicit or explicit models (cf. Rao, 2003, p. 2).

In model-based small area estimation, linear mixed models are commonly applied to exploit similarities between areas and account for the unobserved heterogeneity across areas. They allow for the correlation of observations within areas and yield predictions for the random effects that take the area-specific sample sizes into consideration. The general linear mixed model is described in Section 3.1. Examples of empirical best linear unbiased predictors are the procedures according to Fay and Herriot (1979) and Battese, Harter, and Fuller (1988). While the prediction of point estimates is straightforward in many cases, the assessment of the prediction error can be very cumbersome. Approximations to the MSE under linear mixed models were derived by Prasad and Rao (1990) and Datta and Lahiri (2000). These developments are presented in Section 3.2.

Over the last few years, applications in fields such as poverty measurement required the development of new small area estimation methods suitable for non-linear statistics. The most prominent example in this regard is the empirical best prediction approach from Molina and Rao (2010), which is introduced in Section 3.3

Moreover, even if the desired statistic is linear in the observations, meeting the model assumptions may require to model a one-to-one transformation of the observations instead. The issue of how to obtain small domain estimates in this case is reviewed in Section 3.4.

A key question for the successful implementation of model-based estimation techniques is whether the assumed model can be validated for the data at hand. This topic is explored in Section 3.5, where procedures for model diagnostics and tools for model selection are presented.

The key findings and further points worth discussing when applying model-based small area estimation techniques are summarised in Section 3.6.

3.1 The linear mixed model

Due to its importance, we will outline the general linear mixed model in this section. It is given by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{C}\mathbf{v} + \boldsymbol{\varepsilon}, \tag{3.1.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{X}$ and $\mathbf{C}$ are known $n \times p$ and $n \times q$ matrices, $\mathbf{v}$ is the $q \times 1$ vector of random effects and $\boldsymbol{\varepsilon}$ is a $n \times 1$ error vector (Searle, Casella, & McCulloch, 1992, p. 139). The vector of random effects may be further partitioned into a series of r sub-vectors

$$\mathbf{v} = [\mathbf{v}_1^T \dots \mathbf{v}_r^T]^T$$

(Searle et al., 1992, p. 139). In many applications of model (3.1.1), $\mathbf{C}$ is a matrix that indicates which units of the sample are affected by certain random effects. Following Searle et al. (1992), we may then define

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \mathrm{E}(\mathbf{y}|\mathbf{v}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{C}\mathbf{v}, \tag{3.1.2}$$

where $\mathrm{E}(\cdot)$ is the expectation with respect to model (3.1.1) and can thus express the error vector as

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathrm{E}(\mathbf{y}|\mathbf{v}).$$

Thus, we have to distinguish $E(\mathbf{y}|\mathbf{v})$, which is the conditional mean of $\mathbf{y}$ given the actual random effect, from $E(\mathbf{y})$. The latter is conditional only on the model matrix $\mathbf{X}$ and may be obtained from integrating out the vector of random effects $\mathbf{v}$ as $\mathrm{E}[\mathrm{E}(\mathbf{y}|\mathbf{v})|\mathbf{v}]$. In small area estimation, the conditional mean $\mathrm{E}(\mathbf{y}|\mathbf{v})$ is often the quantity of interest. Since $\mathbf{v}$ is a vector of random effects, we assume $\mathrm{E}(\mathbf{v}) = \mathbf{0}$ that implies together with (3.1.2) $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$. To complete the model, we need the variances of the random effects and the error vector. It is commonly assumed that the $\mathrm{Cov}(\mathbf{v}_i, \mathbf{v}_{i'}^T) = \mathbf{0} \quad \forall i \quad \neq i'$ (cf. Searle et al., 1992, p. 139), i.e. the different random effects are independent from each other, leading to a block-diagonal variance-covariance matrix of $\mathbf{v}$. We may further specify $\mathrm{Var}(\mathbf{v}_i) = \sigma_i^2 \mathbf{I}_{q_i} \forall i$, such that the variance-covariance matrix of $\mathbf{v}$ follows as:

$$\mathrm{Var}(\mathbf{v}) = \bigoplus_{i=1}^{r} \sigma_i^2 \mathbf{I}_{q_i} =: \mathbf{D}, \tag{3.1.3}$$

where $\bigoplus$ denotes the direct sum. Imposing independence of the error term for any two elements in the sample requires the matrix $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{R}$ to be diagonal. Since $\mathbf{v}$ and $\boldsymbol{\varepsilon}$ are the only random parts of model (3.1.1), its variance-covariance matrix is given by

$$\mathbf{V} := \mathrm{Var}(\mathbf{y}) = \mathrm{Var}(\mathbf{C}\mathbf{v}) + \mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{C}\mathbf{D}\mathbf{C}^T + \mathbf{R}. \tag{3.1.4}$$

3.1.1 Prediction of mixed effects

In many applications of general linear mixed models of type (3.1.1) such as small area estimation, the prediction of random variables is of major interest (cf. Jiang & Lahiri, 2006b, p. 9). These predictions may be desired for the random effects $\mathbf{v}$ themselves or for mixed effects of the kind

$$\boldsymbol{\eta} = \mathbf{l}^T\boldsymbol{\beta} + \mathbf{m}^T\mathbf{v}, \tag{3.1.5}$$

where $\mathbf{l}$ and $\mathbf{m}$ are known vectors (cf. Jiang & Lahiri, 2006b, Section 3.2 or Rao, 2003, Section 6.2.1). These quantities can be predicted among other methods using best prediction or best linear unbiased prediction approaches (cf. Rao, 2003, Chapter 6 or Searle et al., 1992, Chapter 7) under a frequentist framework. Additionally, Bayesian methods (cf. Rao, 2003, Chapter 10) may be employed. The best predictor (BP) minimises the MSE of any quantity to be predicted and is given by the conditional expectation given the data (Searle et al., 1992, Section 7.2). Thus, the BP of the vector of random effects $\mathbf{v}$ given $\mathbf{y}$ is simply

$$\mathbf{v}^{BP} := \mathrm{E}(\mathbf{v}|\mathbf{y}) = \int \mathbf{v} f(\mathbf{v}|\mathbf{y}) \mathrm{d}\mathbf{v},$$

which requires knowledge of the conditional distribution of $f(\mathbf{v}|\mathbf{y})$ (cf. Searle et al., 1992, Section 7.2). Note that the best prediction approach yields an unbiased predictor for the vector of random effects under a general linear mixed model of type (3.1.1), as shown by Searle et al. (1992, p. 262). They consider the expectation of the BP over $\mathbf{y}$ and use the law of iterated expectations to get:

$$\mathrm{E}_{\mathbf{y}}(\mathbf{v}^{BP}) = \mathrm{E}_{\mathbf{y}}\left[\mathrm{E}_{\mathbf{v}|\mathbf{y}}(\mathbf{v}|\mathbf{y})\right] = \mathrm{E}(\mathbf{v}).$$

This model-unbiasedness of the BP does not necessarily hold for other models, as we will see later. Searle et al. (1992, Section 7.2) further discuss that the definition of unbiasedness as $\mathrm{E}(\widehat{\mathbf{v}}) = \mathrm{E}(\mathbf{v})$ is different from the usual concept of unbiasedness for a fixed quantity, which implies $\mathrm{E}(\widehat{m}) = m$ where m denotes a constant. If we are willing to assume that $\mathbf{v}$ and $\boldsymbol{\varepsilon}$ are jointly normally distributed, it follows directly from (3.1.1) that $\mathbf{y}$ is marginally multivariate normally distributed as well (cf. Jiang & Lahiri, 2006b, Section 3):

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}). \tag{3.1.6}$$

Moreover, in this case the joint distribution of $\mathbf{v}$ and $\mathbf{y}$ is multivariate normal and given by:

$$\begin{pmatrix} \mathbf{v} \\ \mathbf{y} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{X}\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{D}\mathbf{C}^T \\ \mathbf{C}\mathbf{D} & \mathbf{V} \end{pmatrix}\right). \tag{3.1.7}$$

Further, by virtue of the multivariate normal distribution the conditional distributions are multivariate normal as described in Searle et al. (1992, Appendix S.3) with the following explicit expressions:

$$\mathbf{v}|\mathbf{y} \sim N\left(\mathbf{D}\mathbf{C}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \mathbf{D}-\mathbf{D}\mathbf{C}^T\mathbf{V}^{-1}\mathbf{C}\mathbf{D}\right) \tag{3.1.8}$$

$$\mathbf{y}|\mathbf{v} \sim N\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{C}\mathbf{v}, \mathbf{R}\right). \tag{3.1.9}$$

While (3.1.6) is called a marginal model, (3.1.9) is referred to as a conditional model.

The preceding development is highly relevant if we are interested in deriving the BP of the mixed effect $\boldsymbol{\eta}$ as defined in (3.1.5). Since $\mathbf{v}$ cannot be observed (cf. Jiang & Lahiri, 2006b, p. 9) and is a random vector, the main issue is its prediction given the data at hand. From equation (3.1.8) it is known that

$$\mathrm{E}(\mathbf{v}|\mathbf{y}) = \mathbf{D}\mathbf{C}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}). \tag{3.1.10}$$

Hence, with the normality assumption the problem of predicting the random variable $\mathbf{v}$ is reduced to the estimation of its conditional mean, which is given by (3.1.10) (Searle et al., 1992, p. 262). Furthermore, assuming normality yields an expression for the prediction variance of the conditional mean as $\mathbf{D}-\mathbf{D}\mathbf{C}^T\mathbf{V}^{-1}\mathbf{C}\mathbf{D}$. Substituting the BP (3.1.10) of the random effect into (3.1.5) yields the BP for the mixed effect as (cf. Jiang, 2007, Section 2.3):

$$\widetilde{\boldsymbol{\eta}}^{BP} = \mathbf{l}^T\boldsymbol{\beta} + \mathbf{m}^T\mathbf{D}\mathbf{C}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}). \tag{3.1.11}$$

Under model (3.1.1), the BP of the random effect is model-unbiased and under the assumption of normality of $\boldsymbol{\varepsilon}$ and $\mathbf{v}$ also linear in $\mathbf{y}$. If the normality assumption may seem to be doubtful and a linear predictor is desired, one alternative is to choose the best linear predictor. This approach selects the linear predictor of $\mathbf{v}$ of the type $\widehat{\mathbf{v}} = \mathbf{a}+\mathbf{B}\mathbf{y}$ for known vector $\mathbf{a}$ and matrix $\mathbf{B}$ that minimises the MSE. It can be shown that for model (3.1.1) the best linear predictor is identical to the BP (3.1.10), but without assuming normality (Searle et al., 1992, Section 7.3). Moreover, we might not only require linearity of a predictor but also model-unbiasedness. A predictor fulfilling these constraints will be called a best linear unbiased predictor (BLUP). The following derivation of the BLUP for model (3.1.1) is based on Rao (2003, Section 6.4.1). By $\widetilde{\boldsymbol{\beta}}$, we denote the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ assuming known variance components, given by

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}. \tag{3.1.12}$$

The quantity to be predicted is given by (3.1.5), where $(\mathbf{l}^T,\ \mathbf{m}^T) \neq \mathbf{0}$. The desired predictor is

$$\text{linear in } \mathbf{y}: \quad \mathbf{a}^T\mathbf{y}+b, \quad \mathbf{a}^T \neq \mathbf{0} \tag{3.1.13}$$

$$\text{unbiased:} \quad \mathrm{E}\left(\mathbf{a}^T\mathbf{y}+b-(\mathbf{l}^T\boldsymbol{\beta}+\mathbf{m}^T\mathbf{v})\right) = \mathbf{0}. \tag{3.1.14}$$

As noted by Rao (2003), the condition (3.1.14) implies

$$\mathbf{a}^T\mathbf{X} = \mathbf{l}^T \text{ and } b = 0$$

since the expectation is a linear operator and $\mathrm{E}(\mathbf{v}) = \mathbf{0}$. To obtain the BLUP for the general linear mixed model, we have to minimise the MSE under constraints (3.1.13) and (3.1.14). Since the predictor is unbiased, the MSE is identical to the prediction variance, which is obtained from developing the square of expression (3.1.14). This yields:

$$\mathrm{Var}\left(\mathbf{a}^T\mathbf{y} + b - (\mathbf{l}^T\boldsymbol{\beta} + \mathbf{m}^T\mathbf{v})\right) = \mathbf{a}^T\mathbf{V}\mathbf{a} - 2\mathbf{a}^T\mathbf{C}\mathbf{D}\mathbf{m} + \mathbf{m}^T\mathbf{D}\mathbf{m}.$$

Incorporating the constraint $\mathbf{a}^T\mathbf{X} = \mathbf{l}^T$ by means of the Lagrange multiplier technique yields the following optimisation problem

$$\underset{\mathbf{a}^T,\boldsymbol{\lambda}}{\arg\min} \quad \mathbf{a}^T\mathbf{V}\mathbf{a} - 2\mathbf{a}^T\mathbf{C}\mathbf{D}\mathbf{m} + \mathbf{m}^T\mathbf{D}\mathbf{m} - 2\boldsymbol{\lambda}(\mathbf{a}^T\mathbf{X} - \mathbf{l}^T). \tag{3.1.15}$$

Using the fact that $\mathbf{a}^T\mathbf{V}\mathbf{a}$ and $\mathbf{m}^T\mathbf{D}\mathbf{m}$ are quadratic forms and $\mathbf{D}$ and $\mathbf{V}$ are both symmetric matrices we obtain the first-order conditions as:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{a}} &= 2\mathbf{V}\mathbf{a} - 2\mathbf{C}\mathbf{D}\mathbf{m} + 2\mathbf{X}\boldsymbol{\lambda} \overset{!}{=} \mathbf{0} \\
&\Rightarrow \mathbf{a} = -\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\lambda} + \mathbf{V}^{-1}\mathbf{C}\mathbf{D}\mathbf{m} && (3.1.16)\\
\frac{\partial}{\partial \boldsymbol{\lambda}} &= 2\mathbf{a}^T\mathbf{X} - 2\mathbf{l}^T \overset{!}{=} \mathbf{0} && (3.1.17)\\
&\Rightarrow \mathbf{a}^T\mathbf{X} = \mathbf{l}^T.
\end{aligned}$$

Plugging (3.1.16) into (3.1.17), whose transpose is given by $\mathbf{X}^T\mathbf{a} = \mathbf{l}$ and solving for $\boldsymbol{\lambda}$ yields:

$$\boldsymbol{\lambda} = -(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{l} + (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{C}\mathbf{D}\mathbf{m}.$$

This expression can be used to determine $\mathbf{a}$ as:

$$\begin{aligned}
\mathbf{a} &= -\mathbf{V}^{-1}\mathbf{X}\left[-(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{l} + (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{C}\mathbf{D}\mathbf{m}\right]\mathbf{V}^{-1}\mathbf{C}\mathbf{D}\mathbf{m} \\
&= \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{l} + \left[\mathbf{I}_n - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\right]\mathbf{V}^{-1}\mathbf{C}\mathbf{D}\mathbf{m}.
\end{aligned}$$

Inserting $\mathbf{a}$ into the desired linear predictor gives the BLUP:

$$\begin{aligned}
\widetilde{\boldsymbol{\eta}}^{BLUP} := \mathbf{a}^T\mathbf{y} &= \mathbf{l}^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} + \mathbf{m}^T\mathbf{D}\mathbf{C}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}\right) \\
&= \mathbf{l}^T\widetilde{\boldsymbol{\beta}} + \mathbf{m}^T\mathbf{D}\mathbf{C}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right). && (3.1.18)
\end{aligned}$$

It may be noted that the BLUP (3.1.18) is identical to the BP, except that the latter uses the true $\boldsymbol{\beta}$ whereas the former uses the BLUE of $\boldsymbol{\beta}$ (see Searle et al., 1992, Section 7.4 for a discussion). Furthermore, the assumption of a block-diagonal variance-covariance structure is by no means necessary. Rao (2003, Section 6.2.1) gives a derivation of the BLUP under a general variance-covariance matrix, which yields the same expression as (3.1.18). However, the computation of the inverse of the variance-covariance matrix of $\mathbf{y}$ is tremendously simplified by a block-diagonal $\mathbf{V}$ matrix. Besides this computational aspect, the most commonly used small area models operate under a block-diagonal variance-covariance matrix. Examples with a more complex correlation structure are frequently encountered when considering spatial small area models (cf. Molina, Salvati, & Pratesi, 2009, and references therein).

3.1.2 The MSE of the mixed effect

Up to now, for notational convenience, we simply wrote $\widetilde{\boldsymbol{\eta}}^{BP}$ and $\widetilde{\boldsymbol{\eta}}^{BLUP}$, which suppressed that both predictors are functions of the variance components $\boldsymbol{\delta} = (\sigma_v^2, \sigma_\varepsilon^2)^T$, the data $\mathbf{y}$ and the vector of the regression parameters in the case of the BP or its BLUE in the case of the BLUP. When considering the MSE, however, it will be useful to make this dependencies explicit, as we will see shortly. Thus, we will now indicate the BP as $\widetilde{\boldsymbol{\eta}}^{BP} = \widetilde{\boldsymbol{\eta}}^{BP}(\boldsymbol{\delta}, \mathbf{y}, \boldsymbol{\beta})$ and the BLUP accordingly as $\widetilde{\boldsymbol{\eta}}^{BLUP} = \widetilde{\boldsymbol{\eta}}^{BLUP}(\boldsymbol{\delta}, \mathbf{y}, \widetilde{\boldsymbol{\beta}})$. The MSE of the BP is identical to its prediction variance, since (3.1.11) is an unbiased predictor of (3.1.5). As the only random quantity in the BP is the random effect, using (3.1.8) yields the MSE as (cf. Searle et al., 1992, p. 265)

$$\mathrm{MSE}(\widetilde{\boldsymbol{\eta}}^{BP}(\boldsymbol{\delta}, \mathbf{y}, \boldsymbol{\beta})) = \mathbf{D} - \mathbf{D}\mathbf{C}^T\mathbf{V}^{-1}\mathbf{C}\mathbf{D}. \tag{3.1.19}$$

As outlined by Rao (2003, Section 6.2.2), the BLUP (3.1.18) may be rewritten as

$$\widetilde{\boldsymbol{\eta}}^{BLUP}(\boldsymbol{\delta}, \mathbf{y}, \widetilde{\boldsymbol{\beta}}) = \widetilde{\boldsymbol{\eta}}^{BP}(\boldsymbol{\delta}, \mathbf{y}, \boldsymbol{\beta}) + \mathbf{d}^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

with $\mathbf{d}^T := \mathbf{l}^T - \mathbf{m}^T\mathbf{D}\mathbf{C}^T\mathbf{V}^{-1}\mathbf{X}$. Rao (2003) showed further that the second term of the above equation is uncorrelated with the error of the BP and hence the MSE of the BLUP is given by

$$\begin{aligned}\mathrm{MSE}(\widetilde{\boldsymbol{\eta}}^{BLUP}(\boldsymbol{\delta}, \mathbf{y}, \widetilde{\boldsymbol{\beta}})) &= \mathrm{MSE}(\widetilde{\boldsymbol{\eta}}^{BP}(\boldsymbol{\delta}, \mathbf{y}, \boldsymbol{\beta})) + \mathrm{Var}(\mathbf{d}^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})) \\ &= \mathbf{D} - \mathbf{D}\mathbf{C}^T\mathbf{V}^{-1}\mathbf{C}\mathbf{D} + \mathbf{d}^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{d}.\end{aligned} \tag{3.1.20}$$

The first component of the MSE is due to the prediction of the random effects $\mathbf{v}$, while the second component accounts for the variability due to estimating $\widetilde{\boldsymbol{\beta}}$ (cf. Rao, 2003, Section 6.2.2). Since $\boldsymbol{\beta}$ is assumed to be known in BP, the MSE of the BLUP given by (3.1.20) will be always greater or equal to the MSE of the BP given by (3.1.19).

As pointed out by Jiang (2007, p. 77), neither the fixed effects vector $\boldsymbol{\beta}$ nor the variance components $\boldsymbol{\delta}$ are known in practice, which hence prevents the use of best prediction

and best linear unbiased prediction approaches. The usual strategy is to obtain model-consistent estimates for these parameters and plug these into the different formulae. This procedure is called the empirical best linear unbiased prediction (EBLUP) approach if the BLUP methodology is applied and empirical best prediction (EBP) approach if the BP methodology is used.

Since much of the work in linear mixed models focussed on the EBLUP, we shall focus in the following on the development of it. The EBLUP under model (3.1.1) follows from replacing the unknown variance components $\boldsymbol{\delta}$ by a consistent estimate $\widehat{\boldsymbol{\delta}}$ which is then used to obtain the empirical BLUE $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\widehat{\boldsymbol{\delta}})$ that depends on the vector of variance components. Plugging these quantities into the equation for the BLUP yields the EBLUP according to Rao (2003, Section 6.2.4) as

$$\widehat{\boldsymbol{\eta}}^{EBLUP}(\widehat{\boldsymbol{\delta}}, \mathbf{y}, \widehat{\boldsymbol{\beta}}) := \widehat{\boldsymbol{\eta}}^{BLUP}(\widehat{\boldsymbol{\delta}}, \mathbf{y}, \widehat{\boldsymbol{\beta}}). \tag{3.1.21}$$

Note that the variance components affect the variance-covariance matrices of both $\mathbf{v}$ as well as $\boldsymbol{\varepsilon}$, and hence also the variance-covariance matrix $\mathbf{V}$. The BLUE $\widetilde{\boldsymbol{\beta}}$ given by (3.1.12) in turn depends also on $\mathbf{V}$. Since the estimation of the variance components introduces a further source of uncertainty, this should also be reflected in the MSE. Unfortunately, a closed-form solution for the expression $\mathrm{E}(\widehat{\boldsymbol{\eta}}^{EBLUP}(\widehat{\boldsymbol{\delta}}, \mathbf{y}, \widehat{\boldsymbol{\beta}}) - \boldsymbol{\eta})^2$ cannot be given. Kackar and Harville (1984) proved that if $\widehat{\boldsymbol{\delta}}$ is translation-invariant, which holds for commonly applied estimation such as maximum likelihood (ML), restricted maximum likelihood (REML) or the method of moments (Kackar & Harville, 1981), and normality is assumed, the MSE of (3.1.21) may be decomposed into

$$\begin{aligned}\mathrm{MSE}(\widehat{\boldsymbol{\eta}}^{EBLUP}(\widehat{\boldsymbol{\delta}}, \mathbf{y}, \widehat{\boldsymbol{\beta}})) &= \mathrm{MSE}(\widehat{\boldsymbol{\eta}}^{BLUP}(\boldsymbol{\delta}, \mathbf{y}, \widetilde{\boldsymbol{\beta}})) \\ &\quad + \mathrm{E}\left[(\widehat{\boldsymbol{\eta}}^{EBLUP}(\widehat{\boldsymbol{\delta}}, \mathbf{y}, \widehat{\boldsymbol{\beta}}) - \widetilde{\boldsymbol{\eta}}^{BLUP}(\boldsymbol{\delta}, \mathbf{y}, \widetilde{\boldsymbol{\beta}}))^2\right].\end{aligned} \tag{3.1.22}$$

The second term in (3.1.22), however, does not have a closed-form expression. For some commonly used small area models with block-diagonal variance-covariance matrices, Prasad and Rao (1990) derive an approximation to the expectation of the squared difference between the EBLUP and the BLUP. An extension to the case of the general linear mixed model (3.1.1) was given by Das, Jiang, and Rao (2004).

3.1.3 Estimation of the model parameters

In order to estimate the variance components, consistent procedures such as ML, REML or the method of moments can be employed (cf. Rao, 2003, Section 6.2.4, or Jiang, 2007, Sections 1.3 - 1.5). The log-likelihood under the linear mixed model is given by

$$l(\boldsymbol{\beta}, \boldsymbol{\delta}) = c - \frac{1}{2}\log(|\mathbf{V}|) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{3.1.23}$$

with $|\mathbf{V}|$ as the determinant of the variance-covariance matrix and c as a constant that does not involve any of the model parameters (cf. Jiang, 2007, p. 10). The first order

conditions for the maximum are obtained by differentiating (3.1.23) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ and setting these equations to zero. This yields the well-known expression for $\boldsymbol{\beta}$ (assuming fixed variance components $\boldsymbol{\delta}$) as

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}) \tag{3.1.24}$$

and for the variance components as

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial \delta_r} = \frac{1}{2}\left\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \delta_r}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \delta_r}\right)\right\} = 0,\ r = 1,\ldots,q, \tag{3.1.25}$$

with δ_r as the r-th variance component and $\operatorname{tr}(\cdot)$ as the trace operator. Note that in practice, $\widetilde{\boldsymbol{\beta}}$ cannot be computed and is replace by the empirical BLUE $\widehat{\boldsymbol{\beta}}$. The ML approach yields efficient estimators of the variance components, provided normality is assumed, but suffers from biases, especially with small sample sizes. (cf. Jiang & Lahiri, 2006b, p. 10). An alternative that reduces the bias is the REML approach, where a transformation is applied on the $\mathbf{y}$-vector entering the likelihood equation (3.1.23). The transformation employed is $\mathbf{A}^T\mathbf{y}$, where $\mathbf{A}$ is a $n \times (n-p)$ matrix of full rank that is orthogonal to $\mathbf{X}$. This procedure effectively removes the fixed effects from the likelihood, thereby accounting for the loss in degrees of freedom due to estimating $\boldsymbol{\beta}$ (Rao, 2003, Section 6.2.4). Afterwards, $\boldsymbol{\beta}$ can be estimated from the expression (3.1.24). Note that both ML and REML depend on the assumptions of normality.

Alternatively, if normality cannot be assumed, other methods such as the analysis of variances (ANOVA) method or Henderson's methods, which extend the ANOVA approach to unbalanced data, can be used. They are described in much detail in Searle et al. (1992, Chapter 5). In these approaches, the empirical sum of squared residuals for the variance components is related to their expected values. The estimates of $\widehat{\boldsymbol{\delta}}$ are obtained by the solutions of these equations (cf. Jiang, 2007, Section 1.5.1). Advantages of these approaches are that they do not require normality assumptions of $\boldsymbol{\delta}$ and that these methods are easy to implement. Whilst the solutions to the moments equations are unbiased they are not necessarily part of the admissible parameter space, i.e. negative values may occur for $\delta_r,\ r > 1$. In practice, negative estimates of the variance components are replaced by zero which leads to the problem that the truncated estimates are no longer unbiased.

3.2 Small area estimation under linear mixed models

3.2.1 Nested error regression model

The nested error regression model is a special case of the general linear mixed model and given by (cf. Rao, 2003, Section 7.2.1):

$$
\begin{aligned}
y_{dj} &= \mathbf{x}_{dj}^T\boldsymbol{\beta} + v_d + \varepsilon_{dj}, \quad j = 1, \ldots, n_d, \quad d = 1, \ldots, D, \\
\varepsilon_{dj} &\overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2), \\
v_d &\overset{i.i.d.}{\sim} N(0, \sigma_v^2),
\end{aligned} \tag{3.2.1}
$$

with v_d as area-specific random intercepts with zero mean and variance σ_v^2. Furthermore, independence between v_d and ε_{dj} is assumed. Model (3.2.1) comprises the standard regression model when $\sigma_v^2 = 0$. Note that the nested error regression model is a special case of a two-level model, where only the intercepts are assumed to be random. Due to this, model (3.2.1) is also referred to as a random intercept model. In the small area context, model (3.2.1) is also known as the Battese-Harter-Fuller (BHF) model, since they were the first to apply this model for the prediction of county crop areas in Iowa (Battese et al., 1988).

Now, suppose our aim is to predict the unknown small domain means

$$
\mu_d = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj}, \quad d = 1, \ldots, D, \tag{3.2.2}
$$

where N_d indicates the population size in domain d. These domain means are unknown, since they depend on y_{dj} for all units in each domain, whereas we observe y_{dj} only for the sampled elements $j \in S_d$. Assuming that the nested error regression model (3.2.1) holds for both the population and the sample, we may obtain the BP or the BLUP for the domain mean. The BP of (3.2.2) under model (3.2.1) follows as:

$$
\widetilde{\mu}_d^{BP} = \frac{1}{N_d}\left[\sum_{j\in S_d} y_{dj} + \sum_{j\in U_d\setminus S_d} \widetilde{y}_{dj}^{BP}\right], \quad d = 1, \ldots, D, \tag{3.2.3}
$$

with $\widetilde{y}_{dj}^{BP}$ as the BP for the non-sampled units. The $\widetilde{y}_{dj}^{BP}$ are given by

$$
\begin{aligned}
\widetilde{y}_{dj}^{BP} &= \mathbf{x}_{dj}^T\boldsymbol{\beta} + \widetilde{v}_d, \quad \text{where} \\
\widetilde{v}_d &= \gamma_d(\overline{y}_d - \overline{\mathbf{x}}_d^T\boldsymbol{\beta}).
\end{aligned} \tag{3.2.4}
$$

In (3.2.4), $\overline{y}_d = n_d^{-1}\sum_{j=1}^{n_d} y_{dj}$ and $\overline{\mathbf{x}}_d = n_d^{-1}\sum_{j=1}^{n_d} \mathbf{x}_{dj}$ are the sample means of the dependent variable and the auxiliary information in domain d. Further,

$$
\gamma_d = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\varepsilon^2/n_d}, \tag{3.2.5}
$$

is the shrinkage factor, depending on the variance components and the domain-specific sample size. In practice, the BP (3.2.3) cannot be computed as the true regression vector $\boldsymbol{\beta}$ and the true vector of variance components $\boldsymbol{\delta} = (\sigma_v^2, \sigma_\varepsilon^2)^T$ are not known and have to

be estimated from the data. Replacing $\boldsymbol{\beta}$ by the empirical BLUE $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\delta}$ by a consistent estimate $\widehat{\boldsymbol{\delta}}$ in (3.2.3), leads to the EBP (cf. Rao, 2003, Section 9.3). It can be seen from (3.2.3) that the BP is composed of two parts: the sample information on the dependent variable and the predictions for the non-sampled units given the model (3.2.1). Note that this approach requires the sampling design to be non-informative, such that model (3.2.1) also holds for the population, as pointed out by Rao (2003, Section 5.3). Alternatives that are applicable when the sample is informative will be discussed in Chapter 4.

Alternatively, we may seek the BLUP under model (3.2.1), which is given by:

$$\begin{aligned} \widetilde{\mu}_d^{BLUP} &= \frac{1}{N_d}\left[\sum_{j\in S_d} y_{dj} + \sum_{j\in U_d\setminus S_d} \widetilde{y}_{dj}^{BLUP}\right], \quad d=1,\ldots,D, \quad \text{where} \\ \widetilde{y}_{dj}^{BLUP} &= \mathbf{x}_{dj}^T\widetilde{\boldsymbol{\beta}} + \gamma_d(\overline{y}_d - \overline{\mathbf{x}}_d^T\widetilde{\boldsymbol{\beta}}). \end{aligned} \tag{3.2.6}$$

The EBLUP is obtained from (3.2.6) by replacing the unknown variance components by consistent estimates and using these estimates in the computation of the empirical BLUE $\widehat{\boldsymbol{\beta}}$. Note that the EBLUP, which is identical to the EBP under (3.2.1) (Rao, 2003, Section 9.3), does not require the normality assumptions. To avoid writing the laborious expression EBLUP/EBP all the time, we shall refer to this estimator under model (3.2.1) from now on simply as the BHF estimator. Furthermore, the computation of the BHF estimator can be simplified, since the predictions $\widehat{y}_{dj}^{BHF}$ enter linearly, hence (Rao, 2003, Section 7.2)

$$\widehat{\mu}_d^{BHF} = f_d\overline{y}_d + (1-f_d)(\overline{\mathbf{X}}_{dr}\widehat{\boldsymbol{\beta}} + \widehat{\gamma}_d(\overline{y}_d - \overline{\mathbf{x}}_d^T\widehat{\boldsymbol{\beta}})), \tag{3.2.7}$$

with $f_d = n_d/N_d$ as the sampling fraction in domain d and $\overline{\mathbf{X}}_{dr}$ as the population mean of the auxiliary information of the non-sampled units in domain d. As pointed out by Rao (2003, Section 7.2), $\overline{\mathbf{X}}_{dr} = (N_d\overline{\mathbf{X}}_d - n_d\overline{\mathbf{x}}_d)/(N_d - n_d)$, so that we do not need to know the auxiliary information for all units in the population. It is sufficient to know the population means $\overline{\mathbf{X}}_d$. Furthermore, if the sampling fractions are small, we may approximate the area mean under the nested error regression model as

$$\mu_d \approx \mathrm{E}(\mu_d|v_d) = \overline{\mathbf{X}}_d^T\boldsymbol{\beta} + v_d$$

as the expectation of the mean of ε_{dj} tends to zero (Pfeffermann, 2013). If this approximation is valid, the BHF estimator may be computed as (Rao, 2003, Section 7.2)

$$\begin{aligned} \widehat{\mu}_d^{BHF} &= \overline{\mathbf{X}}_d^T\widehat{\boldsymbol{\beta}} + \widehat{v}_d \\ &= \overline{\mathbf{X}}_d^T\widehat{\boldsymbol{\beta}} + \widehat{\gamma}_d(\overline{y}_d - \overline{\mathbf{x}}_d^T\widehat{\boldsymbol{\beta}}) \\ &= \widehat{\gamma}_d(\overline{y}_d + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_d)^T\widehat{\boldsymbol{\beta}}) + (1-\widehat{\gamma}_d)\overline{\mathbf{X}}_d^T\widehat{\boldsymbol{\beta}}. \end{aligned} \tag{3.2.8}$$

Thus, the EBLUP under model (3.2.1) reduces to a composite estimator consisting of two components: the survey regression estimator $\overline{y}_d + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_d)^T\widehat{\boldsymbol{\beta}}$ and the synthetic regression estimator $\overline{\mathbf{X}}_d^T\widehat{\boldsymbol{\beta}}$. These two components enter the EBLUP as a convex combination with weights $\widehat{\gamma}_d$. A desirable property of the EBLUP (3.2.8) is that more weight is attached to the survey regression estimator as the sample size n_d increases, since $\partial\widehat{\gamma}_d/\partial n_d > 0$. This is in line with the idea that we should be more confident in the direct part as we have more

and more sample information. Furthermore, the predicted random effects are shrunken towards the mean, where the amount of shrinkage depends on the variance components and the sample size. Thus, the shrinkage is smaller if the sample size is large, which may be deemed desirable.

As a side note, we can express (3.2.8) in terms familiar from the econometrics literature. Rearranging the last line in (3.2.8) yields:

$$\widehat{\mu}_d^{BHF} = \widehat{\gamma}_d(\overline{\mathbf{X}}_d^T\widehat{\boldsymbol{\beta}} + (\overline{y}_d - \overline{\mathbf{x}}_d^T\widehat{\boldsymbol{\beta}})) + (1-\widehat{\gamma}_d)\overline{\mathbf{X}}_d^T\widehat{\boldsymbol{\beta}}. \tag{3.2.9}$$

Suppose the sample at hand is model-balanced on the auxiliary information, i.e. $\overline{\mathbf{x}}_d = \overline{\mathbf{X}}_d$. It follows that the first component can be expressed as $\widehat{\gamma}_d\overline{y}_d$. Note that $\overline{y}_d$ is the conditional expectation of the domain mean using a fixed effects approach to estimate the intercepts v_d, where $\widehat{\boldsymbol{\beta}}$ is obtained from fitting model (3.2.1). Under model-balanced sampling the second part in (3.2.9) can be rearranged to $(1-\widehat{\gamma}_d)\overline{\mathbf{x}}_d^T\widehat{\boldsymbol{\beta}}$. It should be noted that $\overline{\mathbf{x}}_d^T\widehat{\boldsymbol{\beta}}$ corresponds to the conditional expectation of the area mean under a complete pooling model, using the regression parameters from model (3.2.1), where no unobserved heterogeneity between the areas is accounted for. Hence, assuming a balanced sample, $\widehat{\mu}_d^{BHF}$ can be expressed as

$$\widehat{\mu}_d^{BHF,Balanced} = \widehat{\gamma}_d\overline{y}_d + (1-\widehat{\gamma}_d)\overline{\mathbf{x}}_d^T\widehat{\boldsymbol{\beta}}. \tag{3.2.10}$$

Thus it can be viewed as a weighted average between the expectation under a fixed effects and a complete pooling model. In this respect, the predictor (3.2.10) attaches more weight to the fixed effects part for large domain-specific sample sizes. As already mentioned above, this is reasonable since the estimation of fixed effects with scarce information is questionable.

To assess the uncertainty of (3.2.8), its MSE is generally used. Assuming normality, Prasad and Rao (1990) decomposed this prediction error into a sum of three components:

$$\mathrm{MSE}(\widehat{\mu}_d^{BHF}) = g_{1d}(\boldsymbol{\delta}) + g_{2d}(\boldsymbol{\delta}) + g_{3d}(\boldsymbol{\delta}). \tag{3.2.11}$$

In (3.2.11), we have:

$$\begin{aligned} g_{1d}(\boldsymbol{\delta}) &= (1-\gamma_d)\sigma_v^2 \\ g_{2d}(\boldsymbol{\delta}) &= (\overline{\mathbf{X}}_d - \gamma_d\overline{\mathbf{x}}_d)^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}(\overline{\mathbf{X}}_d - \gamma_d\overline{\mathbf{x}}_d) \\ g_{3d}(\boldsymbol{\delta}) &= \mathrm{E}[(\widehat{\mu}_d^{BHF} - \mu_d^{BHF})^2], \end{aligned} \tag{3.2.12}$$

where in the last line, the expectation is with respect to the model (3.2.1) and μ_d^{BHF} is the true area mean under this model. The first component $g_{1d}(\boldsymbol{\delta})$ is due to the prediction of the random effect and the second component $g_{2d}(\boldsymbol{\delta})$ is due to the estimation of $\boldsymbol{\beta}$. The third component $g_{3d}(\boldsymbol{\delta})$ is due to the estimation of the variance components and depends on the estimation method used. Unfortunately, the equation (3.2.11) is not really applicable, since knowledge of the variance components is assumed, which does not hold in reality. Replacing the true variance components by consistent estimates $\widehat{\boldsymbol{\delta}}$, Prasad and Rao (1990)

showed that $g_{2d}(\widehat{\boldsymbol{\delta}})$ and $g_{3d}(\widehat{\boldsymbol{\delta}})$ are approximately unbiased, but $\mathrm{E}[g_{1d}(\widehat{\boldsymbol{\delta}})] \approx g_{1d}(\widehat{\boldsymbol{\delta}}) - g_{3d}(\widehat{\boldsymbol{\delta}})$. Hence, a second-order correct estimator of the MSE follows as

$$\widehat{\mathrm{MSE}}(\widehat{\mu}_d^{BHF}) \approx g_{1d}(\widehat{\boldsymbol{\delta}}) + g_{2d}(\widehat{\boldsymbol{\delta}}) + 2g_{3d}(\widehat{\boldsymbol{\delta}}). \tag{3.2.13}$$

Assuming that the variance components are estimated using the ML or REML approach, Datta and Lahiri (2000) derive an expression for the third component as:

$$g_{3d}(\widehat{\boldsymbol{\delta}}) = \frac{n_d^{-2}}{(n_d\widehat{\sigma}_v^2 + \widehat{\sigma}_\varepsilon^2)^3} \left\{\widehat{\sigma}_v^4 I^{\varepsilon\varepsilon} + \widehat{\sigma}_\varepsilon^4 I^{vv} - 2\widehat{\sigma}_v^2\widehat{\sigma}_\varepsilon^2 I^{v\varepsilon}\right\}. \tag{3.2.14}$$

For $g_{3d}(\widehat{\boldsymbol{\delta}})$ the elements $I^{\varepsilon\varepsilon}$, I^{vv} and $I^{v\varepsilon}$ are obtained from the inverse information matrix and their values are:

$$\begin{aligned}
I^{\varepsilon\varepsilon} &= 2\hat{a}^{-1}\sum_{d=1}^{D} n_d^2\hat{\eta}_d^{-2} \\
I^{vv} &= 2\hat{a}^{-1}\sum_{d=1}^{D}\left((n_d-1)\hat{\sigma}_\varepsilon^{-4} + \hat{\eta}_d^{-2}\right) \\
I^{v\varepsilon} &= -2\hat{a}^{-1}\sum_{d=1}^{D} n_d\hat{\eta}_d^{-2} \quad \text{with} \\
\hat{a} &= \left[\sum_{d=1}^{D} n_d^2\hat{\eta}_d^{-2}\right]\left[\sum_{d=1}^{D}\left((n_d-1)\hat{\sigma}_\varepsilon^{-4} + \hat{\eta}_d^{-2}\right)\right] - \left(\sum_{d=1}^{D} n_d\hat{\eta}_d^{-2}\right)^2 \quad \text{and} \\
\hat{\eta}_d &= n_d\hat{\sigma}_v^2 + \hat{\sigma}_\varepsilon^2.
\end{aligned} \tag{3.2.15}$$

For finite populations, an estimator of the MSE was obtained by Prasad and Rao (1990). It is given by

$$\widehat{\mathrm{MSE}}(\widehat{\mu}_d^{EBLUP}) \approx (1-f_d)^2\left(\widehat{\mathrm{MSE}}(\widehat{\mu}_d^{*BHF}) + N_d^{-1}(1-f_d)^{-1}\widehat{\sigma}_\varepsilon^2\right), \tag{3.2.16}$$

where $\widehat{\mathrm{MSE}}(\widehat{\mu}_d^{*BHF})$ uses (3.2.13), replacing $\overline{\mathbf{X}}_d$ by $\overline{\mathbf{X}}_{dr}$, the mean of the non-sampled units.

3.2.2 Area level model

While the nested error regression model introduced in the previous section is a powerful tool, relevant information may be (partially or fully) available at an aggregate level only such that alternative modelling strategies have to be considered. Suppose for now that the auxiliary information is available on the same level on which domain estimates are desired. Furthermore, the covariates consist of the known domain means for the auxiliary information and the dependent variable to be modelled is the vector of the direct means. In this case, we may use area level models, which were pioneered in the small domain context by Fay and Herriot (1979). They used a two-level model, which comprises a

sampling model as the first stage and a linking model as a second stage. Frequently, for the error term normality assumptions are used at both levels. Together this model reads

$$\begin{aligned} &\text{sampling model} \quad && \widehat{\mu}_d^{Dir}|\mu_d \overset{ind}{\sim} N(\mu_d, \psi_d), \quad d = 1, \ldots, D, \\ &\text{linking model} \quad && \mu_d \overset{ind}{\sim} N(\overline{\mathbf{X}}_d^T \boldsymbol{\beta}, \sigma_v^2), \quad d = 1, \ldots, D, \end{aligned} \tag{3.2.17}$$

with $\widehat{\mu}_d^{Dir}$ as the direct estimator, ψ_d as its known variance, μ_d as the true small area mean, $\overline{\mathbf{X}}_d$ as the auxiliary information and σ_v^2 as the error term in the linking model (Jiang & Lahiri, 2006b). Since the two-level model is matched, it can be combined into a linear mixed model as follows (cf. Jiang & Lahiri, 2006b):

$$\begin{aligned} \widehat{\mu}_d^{\text{Dir}} &= \overline{\mathbf{X}}_d^T \boldsymbol{\beta} + v_d + \varepsilon_d, \quad d = 1, \ldots, D \\ \varepsilon_d &\overset{ind}{\sim} N(0, \psi_d) \\ v_d &\overset{i.i.d.}{\sim} N(0, \sigma_v^2). \end{aligned} \tag{3.2.18}$$

Note that unlike in the nested error regression model (3.2.1), we now allow for unequal variances of ε_d. It can be shown (cf. Rao, 2003, Section 7.1) that the BLUP under model (3.2.17) follows as:

$$\begin{aligned} \widetilde{\mu}_d^{\text{FH}} &= \gamma_d \widehat{\mu}_d^{\text{Dir}} + (1 - \gamma_d) \overline{\mathbf{X}}_d^T \widetilde{\boldsymbol{\beta}}, \quad d = 1, \ldots, D \quad \text{where} \\ \gamma_d &= \frac{\sigma_v^2}{\psi_d + \sigma_v^2}. \end{aligned} \tag{3.2.19}$$

In (3.2.19), $\widetilde{\boldsymbol{\beta}}$ denotes the generalised least squares estimator of $\boldsymbol{\beta}$ given by:

$$\widetilde{\boldsymbol{\beta}}(\sigma_v^2) = \left(\frac{\sum_{d=1}^D \overline{\mathbf{X}}_d \overline{\mathbf{X}}_d^T}{\psi_d + \sigma_v^2} \right)^{-1} \frac{\sum_{d=1}^D \overline{\mathbf{X}}_d \widehat{\mu}_d^{\text{Dir}}}{\psi_d + \sigma_v^2}. \tag{3.2.20}$$

As can be seen from equation (3.2.19), the BLUP is a convex combination of the direct estimator and the regression-synthetic component, with weights γ_d and $1 - \gamma_d$ attached. The BLUP puts more weight on the direct component, if the sampling variance ψ_d is small relative to the model variance σ_v^2. Since the direct estimator is unbiased but may suffer from large variances when the sample sizes are small, this is a reasonable weighting procedure. On the other hand, a small model variance implies hardly any variation of domain means after accounting for the auxiliary information $\overline{\mathbf{X}}_d$ by (3.2.17). Hence, the regression-synthetic component will be a good predictor of the domain mean and receive the most weight. Furthermore, as pointed out by Rao (2003, Section 7.1.1), $\widetilde{\mu}_d^{\text{FH}}$ can be applied to general sampling designs, since the design-based direct estimator is modelled using known auxiliary information. Besides, $\widetilde{\mu}_d^{\text{FH}}$ is also design-consistent. As the sample size $n_d \to \infty$, the sampling variance ψ_d approaches zero and the BLUP (3.2.19) reduces to the direct estimator, which is design-unbiased.

To apply the predictor (3.2.19) in practice, we note that σ_v^2, which affects the weights γ_d as well as $\widetilde{\boldsymbol{\beta}}$, is unknown. Replacing σ_v^2 by a consistent estimator $\widehat{\sigma}_v^2$ and consequently also $\widetilde{\boldsymbol{\beta}}$ by $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\widehat{\sigma}_v^2)$ yields the EBLUP

$$\begin{aligned} \widehat{\mu}_d^{\text{FH}} &= \widehat{\gamma}_d \widehat{\mu}_d^{\text{Dir}} + (1 - \widehat{\gamma}_d) \overline{\mathbf{X}}_d^T \widehat{\boldsymbol{\beta}}, \quad d = 1, \ldots, D \quad \text{where} \\ \widehat{\gamma}_d &= \frac{\widehat{\sigma}_v^2}{\psi_d + \widehat{\sigma}_v^2}. \end{aligned} \tag{3.2.21}$$

Similar to the nested error regression model, the EBLUP does not depend on normality assumptions of the random effect or the sampling error (cf. Rao, 2003, Section 5.2). If these assumptions are made, (3.2.21) also coincides with the EBP under model (3.2.18). Moreover, the assumption of known sampling variances may not be fulfilled in reality. In this case, stabilising the estimated variances $\widehat{\psi}_d$ may be achieved by generalised variance function modelling (cf. Rao, 2003, Section 5.2 and Jiang & Lahiri, 2006b, Section 2.1).

To estimate the random effects variance σ_v^2, various methods can be employed, see Datta, Rao, and Smith (2005) for details. In their path-breaking paper, Fay and Herriot (1979) proposed a method of moments estimator based on the residual sum of squares using a weighted least squares approach. Prasad and Rao (1990) suggested a similar method of moments estimator using ordinary least squares, which does not require an iterative procedure. Both approaches have in common that normality assumptions are not required (cf. Rao, 2003, Section 7.1.2). However, the approach of Fay and Herriot is asymptotically more efficient. A drawback of both these approaches is that the solution may be outside the parameter space (Jiang & Lahiri, 2006b, Section 3.1). Alternatively, assuming normality Datta and Lahiri (2000) considered ML and REML approaches to estimate σ_v^2. It can be shown that the resulting estimates yield lower asymptotic variances, but the advantage over the Fay-Herriot method of moments estimator may be rather small as argued by (Rao, 2003, Section 7.1.3). It may be further noted that even under a ML or REML approach estimates $\widehat{\sigma}_v^2 = 0$ may occur. As pointed out by Li and Lahiri (2010) this implies that all the level 2 variation is explained by the fixed effects, which is unrealistic since this requires a perfect model. Moreover, for $\widehat{\sigma}_v^2 = 0$, the EBLUP (3.2.21) reduces to the regression-synthetic estimator, which is prone to over-shrinking. Thus, Li and Lahiri (2010) regard the requirement $\widehat{\sigma}_v^2 > 0$ as desirable. To guarantee a positive variance component estimate in the presence of a small of number of areas, they propose adjusted ML / REML methods. In their approach, the likelihood function of σ_v^2 is multiplied by σ_v^2, which ensures $\widehat{\sigma}_v^2 > 0$ and yields model-consistent estimates.

The MSE of (3.2.21) can be derived along the same lines as for the unit level predictor as

$$\mathrm{MSE}(\widehat{\mu}_d^{FH}) = g_{1d}(\sigma_v^2) + g_{2d}(\sigma_v^2) + g_{3d}(\sigma_v^2) \tag{3.2.22}$$

where (Prasad & Rao, 1990)

$$\begin{aligned} g_{1d}(\sigma_v^2) &= \gamma_d \psi_d \\ g_{2d}(\sigma_v^2) &= (1-\gamma_d)^2 \overline{\mathbf{X}}_d^T \left(\frac{\sum_{d=1}^D \overline{\mathbf{X}}_d \overline{\mathbf{X}}_d^T}{\sigma_v^2 + \psi_d} \right)^{-1} \overline{\mathbf{X}}_d \\ g_{3d}(\sigma_v^2) &= \mathrm{E}[(\widehat{\mu}_d^{FH} - \widetilde{\mu}_d^{FH})^2]. \end{aligned} \tag{3.2.23}$$

An approximation to the MSE estimate was derived by Prasad and Rao (1990) using their method of moments approach to estimate v_d. Assuming normality on v_d and ε_d, this yields the estimated MSE as

$$\widehat{\mathrm{MSE}}(\widehat{\mu}_d^{FH}) \approx g_{1d}(\widehat{\sigma}_v^2) + g_{2d}(\widehat{\sigma}_v^2) + 2g_{3d}(\widehat{\sigma}_v^2). \tag{3.2.24}$$

As noted by Rao (2003, Section 7.1.5), expression (3.2.24) is also valid if σ_v^2 is estimated by REML. The expressions for the g_{3d}-term depend on the estimation procedure used

and are given in Rao (2003, Section 7.1.5), where also MSE approximations valid for the other methods of estimating the variance component are presented. A comparison of the behaviour of the different MSE estimation methods outlined above is given by Datta et al. (2005). Furthermore, P. Lahiri and Rao (1995) show that the MSE estimator due to Prasad and Rao (1990) is robust with respect to non-normal random effects v_d. While the above-mentioned procedures can be used to estimate the unconditional MSE, $\mathrm{E}\left[(\widehat{\mu}_d^{FH} - \mu_d)^2\right]$, in practice the conditional MSE given the small area means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)^T$ may be desired Rao (2003, Section 7.1.6). Datta, Kubokawa, Molina, and Rao (2011) derive an exactly unbiased estimator of this design-based conditional MSE defined as $\mathrm{E}\left[(\widehat{\mu}_d^{FH} - \mu_d)^2 | \boldsymbol{\mu}\right]$. A simulation study conducted by the authors reveals that this estimator can be very unstable when the sampling variance ψ_d is large. In the same paper, a nearly unbiased estimator of the model-based conditional MSE is derived, which conditions on the direct estimator in a particular domain and treats the other direct estimators as random variables (Datta, Kubokawa, et al., 2011). An application of the unconditional and conditional MSE estimators to Spanish labour force survey data indicated only minor differences in most small domains. An exception is the case when large absolute residuals $|\widehat{\mu}_d^{Dir} - \overline{\mathbf{X}}_d^T \widehat{\boldsymbol{\beta}}|$ occur in domains with small sampling variances ψ_d as observed by Datta, Kubokawa, et al. (2011).

3.3 Non-linear empirical best prediction

Molina and Rao (2010) propose a general procedure to approximate the empirical best predictor for non-linear small area parameters. Their procedure is motivated with an application to small domain estimation of poverty measures in mind. Nonetheless it can be used to predict other quantities as well such as domain means or quantiles that can be adequately modelled using a non-linear transformation.

Their approach assumes the availability of unit level information to produce small area estimates of general non-linear parameters, which do not need to be separable

$$\theta_d = h(\mathbf{y}_d), \quad d = 1, \ldots D, \tag{3.3.1}$$

where $\mathbf{y}_d$ denotes the response vector of all elements (sampled and non-sampled) belonging to area d (Molina & Rao, 2010, p. 374). A simplification arises when the target parameter is additive and separable, as is the case for the family of poverty measures due to Foster, Greer, and Thorbecke (1984). They consider the family of poverty measures given by:

$$\mathrm{FGT}_{\alpha,d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \underbrace{\mathbb{I}(y_{dj} < z) \left(\frac{z - y_{dj}}{z} \right)^{\alpha}}_{:=h(y_{dj})}, \quad \alpha \geq 0, \tag{3.3.2}$$

where z is the poverty line and the choice of alpha corresponds to different commonly used poverty statistics, such as the headcount ratio ($\alpha = 0$) or the poverty gap ($\alpha = 1$). When the target parameter is additive and separable, (3.3.1) can be rewritten as

$$\theta_d = \frac{1}{N_d}\left[\sum_{j\in S_d} h\left(y_{dj}\right) + \sum_{j\in U_d\setminus S_d} h\left(y_{dj}\right)\right], \quad d = 1,\ldots,D. \tag{3.3.3}$$

In equation (3.3.3), the first sum within the brackets is over all units which belong to domain d and are part of the sample, $j \in S_d$, whereas the second sum is over the non-sampled part of domain d, $j \in U_d \setminus S_d$. It should be noted that (3.3.3) contains the estimation of domain means of the y_{dj} as a special case, when $h(y_{dj}) = y_{dj}$. In this case, (3.3.3) reduces to (3.2.2). While for the sampled part of equation (3.3.3) $h(y_{dj})$ is known, a model is needed to derive predictions for the non-sampled part of the population. Since the distribution of the y_{dj} may be fairly complex in many applications, such as poverty mapping, modelling them directly can be inconvenient. As noted by Molina and Rao (2010, p. 372), this problem can be simplified provided a one-to-one transformation $\zeta_{dj} = g(y_{dj})$ such that ζ_{dj} can be adequately modelled using the nested error regression model, exists and is known. This yields

$$\begin{aligned} \zeta_{dj} &= \mathbf{x}_{dj}^T\boldsymbol{\beta} + v_d + \varepsilon_{dj}, \\ \varepsilon_{dj} &\overset{i.i.d}{\sim} N(0,\sigma_\varepsilon^2), \\ v_d &\overset{i.i.d}{\sim} N(0,\sigma_v^2). \end{aligned} \tag{3.3.4}$$

If this model holds for both the population and the sample, the vectors $\boldsymbol{\zeta}_d$ obtained from stacking all elements within a particular domain into a column vector are independently distributed (cf. Molina & Rao, 2010, p. 375). This enables the following decomposition into sampled and non-sampled units:

$$\begin{pmatrix}\boldsymbol{\zeta}_{ds}\\ \boldsymbol{\zeta}_{dr}\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{X}_{ds}\boldsymbol{\beta}\\ \mathbf{X}_{dr}\boldsymbol{\beta}\end{pmatrix}, \begin{pmatrix}\mathbf{V}_{dss} & \mathbf{V}_{dsr}\\ \mathbf{V}_{drs} & \mathbf{V}_{drr}\end{pmatrix}\right). \tag{3.3.5}$$

Note that the above decomposition ensures that the conditional distribution of the transformed non-sampled units $\boldsymbol{\zeta}_{dr}$ given the transformed sampled units $\boldsymbol{\zeta}_{ds}$ is normal with

$$\mathrm{E}\left(\boldsymbol{\zeta}_{dr}|\boldsymbol{\zeta}_{ds}\right) = \mathbf{X}_{dr}\boldsymbol{\beta} + \mathbf{V}_{drs}\mathbf{V}_{dss}^{-1}\left(\boldsymbol{\zeta}_{ds} - \mathbf{X}_{ds}\boldsymbol{\beta}\right) \tag{3.3.6}$$

$$\mathrm{Var}\left(\boldsymbol{\zeta}_{dr}|\boldsymbol{\zeta}_{ds}\right) = \mathbf{V}_{drr} - \mathbf{V}_{drs}\mathbf{V}_{dss}^{-1}\mathbf{V}_{dsr}. \tag{3.3.7}$$

Under these circumstances, the MSE-minimal predictor for the non-sampled units of ζ_{dj} is given by (3.3.6). However, to produce the statistic of interest (3.3.3), the function $h(\cdot)$ has to be applied to the y_{dj} -values. Hence, a predictor of (3.3.3) will be of the form

$$\widehat{\theta}_d = \frac{1}{N_d}\left[\sum_{j\in S_d} h\left(y_{dj}\right) + \sum_{j\in U_d\setminus S_d} h\left(g^{-1}\left(\widehat{\zeta}_{dj}\right)\right)\right], \quad d = 1,\ldots,D. \tag{3.3.8}$$

In the case of non-linear transformations, simply applying the back-transformation to the MSE-optimal predictor $\widetilde{\zeta}_{dj}^{BP}$ will not be unbiased for y_{dj} and hence generally, also

$h\left(g^{-1}\left(\widetilde{\zeta}_{dj}\right)\right) \neq h\left(y_{dj}\right)$. Instead, the MSE-optimal predictor of a target parameter (3.3.3) can be expressed as (Molina & Rao, 2010)

$$\widetilde{\theta}_d^{BP} = \frac{1}{N_d}\left[\sum_{j\in S_d} h\left(y_{dj}\right) + \sum_{j\in U_d\setminus S_d} \mathrm{E}\left(h\left(g^{-1}\left(\widetilde{\zeta}_{dj}\right)\right)|\boldsymbol{\zeta}_{ds}\right)\right], \quad d = 1,\ldots,D. \tag{3.3.9}$$

Since the second term of (3.3.9) does not have a closed form in general, Molina and Rao (2010, p. 373) propose to use Monte-Carlo integration to approximate this expectation as

$$\mathrm{E}\left(h\left(g^{-1}\left(\widetilde{\zeta}_{dj}\right)\right)|\boldsymbol{\zeta}_{ds}\right) \approx \frac{1}{R}\sum_{r=1}^{R} h\left(g^{-1}\left(\widehat{\zeta}_{dj}^{(r)}\right)\right), \tag{3.3.10}$$

where $\widehat{\zeta}_{dj}^{(r)}$ denotes the prediction of the response for the r-th data set on the transformed scale using estimated model parameters. To achieve a high accuracy with approximation (3.3.10), the number of Monte-Carlo replications R should be sufficiently high. It has been noted by Molina and Rao (2010) that with additive separable predictors of type (3.3.3), numerical integration methods may be used to approximate $\mathrm{E}\left(h\left(g^{-1}\left(\widetilde{\zeta}_{dj}\right)\right)|\boldsymbol{\zeta}_{ds}\right)$ instead. The virtue of Monte-Carlo integration techniques is that the error in (3.3.10) is independent of the number of dimensions. Hence, this method is especially suitable for the integration of high-dimensional integrals.

It has been pointed out by Pfeffermann (2013) that the approach due to Molina and Rao (2010) amounts to imputing the values $h(y_{dj})$ of the non-sampled units. Specifically for separable parameters, the point estimate obtained by combining (3.3.8) and (3.3.10) coincides with a point estimate obtained by multiple imputation using model (3.3.4) (cf. Little & Rubin, 2002, Section 5.4). However, the number of replications used when computing (3.3.10) is typically much larger than the number of imputations in a multiple imputation approach. Meinfelder (2014) explained this by the fact that the fraction of missing information in small area estimation is much higher than in standard applications of multiple imputation. Moreover, as stated by Lehtonen, Veijanen, Myrskylä, and Valaste (2011, Section 3.7) "any imputation method could be used to impute all values in the unknown part of the population".

3.4 Small area estimation using transformations

In the previous section, we considered the application of a random intercept model to a general one-to-one transformation of the dependent variable. In many instances, this will be required as the support of the dependent variable may be constrained to non-negative values such that modelling the data as normally distributed is questionable. Moreover, a linear relationship may not hold for the raw values, a typical case for business surveys as discussed by Chandra and Chambers (2011). An often encountered characteristic of business data is that few large companies account for the vast majority of the total of the variables of interest. This implies that the underlying distributions are very skewed

and inherently non-normal, giving rise to non-linear transformations. Since linear models are easy to use and powerful tools in statistical modelling, a lot of effort has been made to apply transformations such that after transforming the data, linear models can be reasonably assumed. If, on the other hand, the raw data exhibit a linear relationship which is distorted by influential outliers, robust estimation techniques may be considered. Amongst these methods is the robust EBLUP due to Sinha and Rao (2009). This approach closely resembles the traditional unit level EBLUP (3.2.7). The point of departure is that the likelihood is modified to avoid having influential values changing the shape of the regression surface too much. This is achieved by truncating large absolute residuals.

In this work, we do not cover robust estimation methods in greater detail. Instead we focus on the use of non-linear transformations, which is also one way of dealing with outliers. In particular, we concentrate on the use of a log transformation, which is a commonly used in business surveys before applying linear models (cf. Chandra & Chambers, 2011, p. 39). Since our sole aim is to predict finite population quantities which are a function of the y_{dj}, we will not make any statements about the auxiliary information and take for granted that its components are such that a linear relationship between y_k and $\mathbf{x}_k$ seems plausible. To assume that $\mathbf{y}|\mathbf{x}$ is lognormally distributed requires $y_k > 0\ \forall k$ as $\log(y_k)$ is not defined otherwise. This may pose problems if the dependent variable has some negative values. However, if the number of negative values is not large, one may add a small constant c such that $\log(y_k + c)\forall k \in U$ (Lehtonen et al., 2011, p. 29). Obviously, this constant has to be taken into account when the back-transformation is applied.

Karlberg (2000) derived a model-unbiased predictor under a lognormal fixed effects superpopulation model. Their model is given by

$$\log(y_k)|\mathbf{x}_k \overset{ind}{\sim} N(\mu_k, \sigma^2), \tag{3.4.1}$$

where $\mu_k = \mathbf{x}_k^T\boldsymbol{\beta}$. Since the EBP under model (3.4.1) is non-linear in the model parameters, it will be biased. To overcome this issue, Karlberg (2000) applied a bias correction to the predictions μ_k from the model.

With respect to small area estimation using log transformations it is interesting to note that already Fay and Herriot (1979) modelled the logarithm of a dependent variable in their pioneering paper. However, they did not consider the issue of bias correction after applying the back-transformation. You and Rao (2002b) proposed an area level model in which the sampling and linking model are not matched, i.e. cannot be expressed as a linear mixed model. They assumed the following model

$$\begin{aligned} \text{sampling model:} \quad & y_d = \mu_d + \varepsilon_d, \quad d = 1, \ldots, D, \\ \text{linking model:} \quad & \log(\mu_d) = \overline{\mathbf{X}}_d^T\boldsymbol{\beta} + v_d, \quad d = 1, \ldots, D. \end{aligned} \tag{3.4.2}$$

and used a HB approach with Gibbs sampling to produce estimates $\widehat{\mu}_d$. They applied model (3.4.2) to estimate the undercoverage from the Canadian census in 1991, and showed that the resulting predictions had in general smaller coefficients of variation than the corresponding direct estimates. Maiti and Slud (2002) compared various area level models for small area estimation in the SAIPE project. Their aim is to predict the child poverty rate on county level. They use the log transformation of direct estimates of the poverty

rate as their dependent variable to fulfil the assumptions of the FH model. Thus, the quantity to be estimated under their model is given by

$$\mu_d = \exp\left(\overline{\boldsymbol{X}}_d^T \boldsymbol{\beta} + v_d\right). \tag{3.4.3}$$

A predictor of (3.4.3) which takes into account the bias due to the back-transformation but ignores the variability of the parameter estimates follows as

$$\widehat{\mu}_d^{ALLOG} = \exp\left(\overline{\boldsymbol{X}}_d^T \widehat{\boldsymbol{\beta}} + \widehat{v}_d + 0.5\widehat{\sigma}_v^2\left(1 - \widehat{\gamma}_d\right)\right), \tag{3.4.4}$$

where $\widehat{v}_d = \widehat{\gamma}_d(\log(y_d) - \overline{\boldsymbol{X}}_d^T \widehat{\boldsymbol{\beta}})$ and $\widehat{\gamma}_d = \widehat{\sigma}_v^2/(\widehat{\sigma}_v^2 + \sigma_{\varepsilon,d}^2)$ (Maiti & Slud, 2002). In the preceding equation $\sigma_{\varepsilon,d}^2$ is the variance of the log-transformed direct estimator, which can be approximated by linearisation as $\sigma_{\varepsilon,d}^2 = \psi_d/\widehat{y}_d^2$ (Lehtonen & Pahkinen, 2004, p. 141). To estimate the MSE of predictor (3.4.4), Maiti and Slud (2002) proposed approximating it by a first-order Taylor expansion. An MSE estimator of this kind is not second-order correct as pointed out by Maiti (2004). Instead they propose to apply the jackknife due to Jiang, Lahiri, Wan, et al. (2002) to estimate the MSE of (3.4.4). Furthermore, Slud and Maiti (2006) obtained a closed-form expression for the MSE estimate of (3.4.4), which is valid provided the number of domains is large.

In some cases, unit level data may be available for the dependent variable for all sampled elements and for the auxiliary variables for all elements within the population. Hence, more efficient unit level models may be applied. Berg and Chandra (2014) extended Karlberg's model (3.4.1) to the case of a lognormal mixed model. They considered the following model:

$$\log(y_{dj}) = \boldsymbol{x}_{dj}^T \boldsymbol{\beta} + v_d + \varepsilon_{dj}, \quad d = 1, \ldots, D,\ j = 1, \ldots, N_d, \tag{3.4.5}$$

where $(v_d, \varepsilon_{dj}) \sim N(0, \mathrm{diag}(\sigma_v^2, \sigma_\varepsilon^2))$ (Berg & Chandra, 2014). Obviously, model (3.4.5) is just the well-known nested error regression model with $\log(y_{dj})$ as dependent variable. Similar to Karlberg, Berg and Chandra developed a bias-corrected predictor for the domain mean, given by

$$\widehat{\mu}_d^{ULSyn} = \frac{1}{N_d}\left(\sum_{j \in s_d} y_{dj} + \sum_{j \notin s_d} \widehat{y}_{dj}^{ULSyn}\right) \tag{3.4.6}$$

where

$$\widehat{y}_{dj}^{ULSyn} = \exp\left(\boldsymbol{x}_{dj}^T \widehat{\boldsymbol{\beta}} + 0.5\left(\widehat{\sigma}_v^2 + \widehat{\sigma}_\varepsilon^2 - \boldsymbol{x}_{dj}^T \widehat{\boldsymbol{V}}(\widehat{\boldsymbol{\beta}})\boldsymbol{x}_{dj} - 0.25\widehat{\boldsymbol{V}}(\widehat{\sigma}_v^2 + \widehat{\sigma}_\varepsilon^2)\right)\right). \tag{3.4.7}$$

In (3.4.7) $\widehat{\boldsymbol{V}}(\widehat{\boldsymbol{\beta}})$ denotes the estimated variance-covariance matrix of the fixed effects and $\widehat{\boldsymbol{V}}(\widehat{\sigma}_v^2 + \widehat{\sigma}_\varepsilon^2)$ refers to the estimated variance-covariance matrix of the random effects and

the error term. Note that the latter depends on the estimation method used for the variance components. Expressions are given in Rao (2003, Section 7.2.3). As discussed by Berg and Chandra (2014, Appendix C), predictor (3.4.6) is a synthetic estimator, since the random effects are not included in the predictions $\widehat{y}_{dj}^{ULSyn}$.

Moreover, Chandra and Chambers (2011) proposed a model-based direct estimator under model (3.4.5). In their approach, the small area mean is estimated by a weighted sum of the sampled units, where the weights are model-calibrated. Thus, the sum of the weights reproduces the known population size and the weighted sum of the fitted values in the sample agrees with the sum of the fitted values in the population. The authors further show that their method is robust with respect to the distribution of the random effects (Chandra & Chambers, 2011, p. 44).

Furthermore, Berg and Chandra (2014) derive the optimal predictor under model (3.4.5), noting that the BP minimising the MSE is given by the conditional expectation given the data, i.e $E(\mu_d|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\xi})$ where $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \sigma_v^2, \sigma_\varepsilon^2)^T$ denotes the vector of model parameters. They derive a closed-form expression for the BP under model (3.4.5) as

$$\widetilde{\mu}_d^{BPLog} = \frac{1}{N_d}\left[\sum_{j\in S_d} y_{dj} + \sum_{j\notin S_d} \widetilde{y}_{dj}^{BPLog}\right], \quad d = 1, \ldots, D \quad \text{where} \tag{3.4.8}$$

$$\widetilde{y}_{dj}^{BPLog} = \exp(\mathbf{x}_{dj}^T\boldsymbol{\beta} + \widetilde{v}_d + 0.5\sigma_\varepsilon^2(\gamma_d/n_d + 1)), \tag{3.4.9}$$

where $\gamma_d = \sigma_v^2/(\sigma_v^2 + \sigma_\varepsilon^2/n_d)$ and $\widetilde{v}_d = \gamma_d\left(\bar{l}_d - \overline{\mathbf{x}}_d^T\boldsymbol{\beta}\right)$ with $\bar{l}_d = n_d^{-1}\sum_{j=1}^{n_d}\log(y_{dj})$. It may be noted that this predictor has the same structure as the BP under a unit level linear mixed model. Inside the brackets is the sum of the sampled units within a domain plus the sum of the best predictions given the model and the available data for the non-sampled units. The BP defined by (3.4.8) cannot be computed in practice, as the model parameters $\boldsymbol{\xi}$ are generally not known and have to be replaced by estimates $\widehat{\boldsymbol{\xi}}$. These estimates are obtained from fitting model (3.4.5) to the sample data, which can be done by REML or ML. Berg and Chandra (2014) suggest using REML to fit model (3.4.5) and obtain the EBP:

$$\widehat{\mu}_d^{EBPLog} = \frac{1}{N_d}\left[\sum_{j\in S_d} y_{dj} + \sum_{j\notin S_d} \widehat{y}_{dj}^{EBPLog}\right], \quad d = 1, \ldots, D \quad \text{where} \tag{3.4.10}$$

$$\widehat{y}_{dj}^{EBPLog} = \exp(\mathbf{x}_{dj}^T\widehat{\boldsymbol{\beta}} + \widehat{v}_d + 0.5\widehat{\sigma}_\varepsilon^2(\widehat{\gamma}_d/n_d + 1)). \tag{3.4.11}$$

Interestingly, the EBP due to Berg and Chandra (2014), can be considered a special case of the general approach due to Molina and Rao (2010), where $h(y_{dj}) = y_{dj}$ and $g^{-1}(\widehat{\zeta}_{dj}) = \exp(\widehat{\zeta}_{dj})$. Owing to the characteristics of the lognormal distribution, the EBP of the non-sampled units given the sampled units has the convenient closed-form solution given by (3.4.10), thereby avoiding the need for numerical or Monte-Carlo integration.

Berg and Chandra (2014) remark that the EBP (3.4.10) is not model-unbiased, since the estimates $\widehat{\boldsymbol{\xi}}$ contribute non-linearly to the predictor. To correct for the bias due to the back-transformation, Berg and Chandra (2014) approximate $\widehat{y}_{dj}^{EBPLog}$ by a Taylor series

to obtain a multiplicative correction. Whilst the EBP (3.4.10) is fairly easy to compute, obtaining a measure of accuracy is much more sophisticated. Based on a proposal by S. N. Lahiri, Maiti, Katzoff, and Parsons (2007), Berg and Chandra (2014) obtain a MSE estimator given by

$$\mathrm{MSE}(\widehat{\mu}_d^{EBPLog}) = M_{1d}(\widetilde{\boldsymbol{\xi}}_d) + M_{2d}(\widehat{\boldsymbol{\xi}}), \tag{3.4.12}$$

where the leading term M_{1d} depends on parameters $\widetilde{\boldsymbol{\xi}}_d$, which are perturbed for area d to correct for the bias due to the non-linear contribution of the model parameters for the leading term. As explained by Berg and Chandra (2014), the derivations involved in the M_{2d}-term rely on a Taylor series argument. Another possibility is to apply the jackknife due to Jiang et al. (2002), which also yields a bias correction for the leading term and evaluates the M_{2d}-term conveniently by resampling. Alternatively, we may apply a parametric bootstrap suitable for the estimation of finite population quantities. Such a bootstrap was proposed by González-Manteiga, Lombardía, Molina, Morales, and Santamaría (2008) for the estimation of the MSE of EBLUPs based on linear mixed models. Their bootstrap can also handle non-normal error distributions and was applied by Molina and Rao (2010) to general non-linear EB predictors. Since the EBP (3.4.10) constitutes a special case of the approach due to Molina and Rao (2010), the parametric bootstrap can be applied to estimate the MSE of the EBP as well. Its application requires the steps outlined in Algorithm 1.

Algorithm 1 MSE estimation by parametric bootstrap for the EBP (3.4.10)

1. Fit model (3.4.5) to the sample data to obtain estimates $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\beta}}^T, \widehat{\sigma}_v^2, \widehat{\sigma}_\varepsilon^2)^T$.
2. Generate bootstrap random area effects $v_d^* \overset{i.i.d.}{\sim} N(0, \widehat{\sigma}_v^2), \quad d = 1, \ldots, D$.
3. Generate, independently from v_d^*, the element-specific error terms
$$\varepsilon_{dj}^* \overset{i.i.d.}{\sim} N(0, \widehat{\sigma}_\varepsilon^2), \quad d = 1, \ldots, D, \quad j = 1, \ldots, N_d.$$
4. Construct a bootstrap population as $\log(y_{dj}^*) = \mathbf{x}_{dj}^T \widehat{\boldsymbol{\beta}} + v_d^* + \varepsilon_{dj}^*, \quad d = 1, \ldots, D, \quad j = 1, \ldots, N_d$. Calculate the domain-means for the bootstrap population as:
$$\mu_d^* = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}^*.$$
5. Using the sample indices from the original sample, S, fit (3.4.5) for the sampled elements from the bootstrap populations to obtain $\widehat{\boldsymbol{\xi}}^*$.
6. Compute the EBP-predictor $\widehat{\mu}_d^{EBPLog*}$ using y_{dj}^* and $\widehat{\boldsymbol{\xi}}^*$ according to (3.4.10).
7. Repeat steps 2. - 6. R-times, where R is a large number.

Using the above steps, an MSE estimator can be computed according to

$$\widehat{\mathrm{MSE}}(\widehat{\mu}_d^{EBPLog}) = \frac{1}{R}\sum_{r=1}^{R}(\widehat{\mu}_d^{EBPLog*,r} - \mu_d^{*,r})^2. \tag{3.4.13}$$

As noted by Molina and Rao (2010), the accuracy of the MSE estimator (3.4.13) can be improved by resorting to a double bootstrap procedure as proposed by Hall and Maiti (2006). Applying their approach, however, can be very computationally demanding.

3.5 Model validation

All the techniques presented in this chapter are model-based. Hence, their applicability critically hinges on the validity of the assumed model. In this section we therefore give a brief presentation of some of the tools available to assess the appropriateness of a model. A couple of diagnostics suitable for linear mixed models with a block-diagonal variance-covariance matrix are introduced in Rao (2003, Section 6.3.4). Moreover Jiang (2007, Section 2.4) discusses some diagnostic tools. An overview of more recent developments is given in Pfeffermann (2013, Section 8).

3.5.1 Model diagnostics

Brown, Chambers, Heady, and Heasman (2001) consider various diagnostics applicable to area level models. They propose a bias diagnostic, which is based on the reasoning that the direct estimates, while suffering from large variances, are still unbiased. Hence, if the model-based estimates are unbiased as well, the design-based estimates should randomly fluctuate around the model-based estimates. This can be visually inspected by means of a scatterplot of the design-based estimates against the model-based estimates with a superimposed regression line. Provided unbiasedness, the regression line should be close to the 45 degree line. Another very useful diagnostic put forward by Brown et al. (2001) is with respect to the coherence of the small area estimates. They authors propose to aggregate model-based small area estimates to larger domains determined by content or geography, for which reliable design-based estimates can be obtained. If the relative difference between both set of estimates is not negligible in a particular large domain, this may indicate a spatial pattern not reflected in the model. Furthermore, as discussed by Rao (2003, Section 6.2.3) the linear mixed model can be transformed to a model with independently and identically distributed errors, such that standard analyses of residuals or influential units can be applied. For the latter, Cooks's distance can be extended towards linear mixed models as explained by Rao (2003, Section 6.2.3), which allows to identify influential observations or areas. The analysis of residuals may be used to identify a possible misspecification of the model. This would be indicated by a visible trend in a scatter plot of the transformed residuals against the transformed fitted values. Moreover, to check whether MSE estimation techniques that crucially depend on normality assumptions are applicable, quantile-quantile (QQ) plots of the transformed residuals may be examined, as conducted by Battese et al. (1988). Besides, also the

standardised residuals, which are approximately standard normally distributed (cf. Rao, 2003, Section 7.2.1), may be studied. Longford (2001) introduced diagnostics to identify outlying level two units under the general linear mixed model (3.1.1). They assume that except for one area all other areas are well described by model (3.1.1), while the remaining area can be described by another linear model. To assess whether a unit is an outlier, they propose to compute a test statistic and to compare against the distribution obtained by parametric bootstrap from the fitted model. If the value of the test statistic is not in the tails of the simulated distribution, the area is unlikely to be an outlier.

3.5.2 Model selection

A vast amount of literature exists on procedures for model selection for linear mixed models. A comprehensive account on this issue is given by Müller, Scealy, and Welsh (2013). Since many of the most frequently used models in small area estimation are random intercept models, ideas in the general multilevel modelling literature apply as well. There is one caveat, though, in that the model parameters are not of great interest as such in small domain estimation. Instead the predictive accuracy of the models is of utmost importance. One of the most popular techniques for model selection in statistical applications is the Akaike information criterion (AIC), which can be used to compare models according to their predictive accuracy. The AIC is given by (cf. Vaida & Blanchard, 2005, p. 353)

$$\text{AIC} = -2\log(g(\mathbf{y}|\widehat{\boldsymbol{\xi}}(\mathbf{y}))) + 2K, \tag{3.5.1}$$

where $\log(g(\mathbf{y}|\widehat{\boldsymbol{\xi}}(\mathbf{y})))$ denotes the log-likelihood evaluated at the (RE)ML estimator $\widehat{\boldsymbol{\xi}}(\mathbf{y})$ and K is a term related to model complexity. Thus, the AIC penalizes both a lack of fit as represented by minus twice the log-likelihood but also the model complexity. Vaida and Blanchard (2005) argue that its application to linear mixed models is not straightforward and should depend on the purpose of using the model. They raise the distinction between the marginal model (3.1.6) and the conditional model (3.1.9) mentioned in Section 3.1. In their view, the selection should be based on the marginal AIC (3.5.1), if the aim of the analysis is to make predictions for units in non-sampled areas. This approach has been called the "population focus" by Vaida and Blanchard (2005, p. 362) and is also the usual procedure in analytical inference, where the structure and knowledge about the underlying model parameters $\boldsymbol{\xi}$ is the aim of the study. On the other hand, if predictions are desired for units in an area which is already included in the sample, the marginal AIC is not the right criterion as it amounts to integrating out the random effects. Instead the conditional AIC proposed by Vaida and Blanchard (2005) could be applied. The formula for the conditional AIC is

$$\text{cAIC} = -2\log g(\mathbf{y}|\widehat{\boldsymbol{\xi}}(\mathbf{y}), \widehat{\mathbf{v}}(\mathbf{y})) + 2K, \tag{3.5.2}$$

with $\log g(\mathbf{y}|\widehat{\boldsymbol{\xi}}(\mathbf{y}), \widehat{\mathbf{v}}(\mathbf{y}))$ as the log-likelihood conditional on the model parameters and the predictions of the random effects (Vaida & Blanchard, 2005, p. 355). The precise expression of the model complexity K depends on the method used to fit the model.

Unlike the marginal AIC, (3.5.2) explicitly accounts for the predicted random effects, which is also reflected in the degrees of freedom, which are in between the marginal model and a model with area-specific fixed effects. Hence, a relevant question in this regard is whether the domain effects should be treated as random or as fixed effects (cf. Vaida & Blanchard, 2005, p. 354). In order to simplify computations, Vaida and Blanchard (2005) ignore the uncertainty associated with the estimation of the random effects as pointed out by Greven and Kneib (2010). As shown by Greven and Kneib (2010), this causes the cAIC to favour a model with random effects over a model without unless the estimated random effects happens to be zero. This feature is clearly undesirable, as the model complexity is not penalised in this case. Greven and Kneib (2010) obtain an analytical expression for a corrected cAIC, which takes into account the uncertainty due to predicting the random effects. It should be noted that the cAIC can be used for both the selection of the covariates and the modelling of unobserved heterogeneity after accounting for fixed predictors. Thus, the cAIC is a very powerful tool in small area modelling. In some applications, however, the selection of the covariates may be unambiguous such that the relevant modelling choice is with respect to the presence of a random effect. In this case, the relevant null hypothesis is $H_0 : \sigma_v^2 = 0$ versus $H_1 : \sigma_v^2 > 0$, which can be tested by a likelihood-ratio test (LRT) or a restricted likelihood-ratio test (RLRT). As pointed out by Crainiceanu and Ruppert (2004), the distribution of the (R)LRT under the null hypothesis is non-standard, since σ_v^2 is on the boundary of the parameter space and the responses are not independent under the alternative. Crainiceanu and Ruppert (2004) derive the finite sample distribution of the (R)LRT under the H_0 and provide efficient algorithms to simulate the (R)LRT. Another strategy specifically tailored towards small area estimation has been proposed by Datta, Hall, and Mandal (2011) in the context of the Fay-Herriot model (3.2.18). Their development is based on the idea that if the variation of the small area means can be sufficiently explained by the fixed part of the model, the use of random effects can be avoided altogether (cf. Datta, Hall, & Mandal, 2011). This will be desirable for the precision of the estimates, since the estimators converge at a faster rate as detailed by Datta, Hall, and Mandal (2011, p. 362f.). A parametric bootstrap approach is employed to test for the presence of a random effect (cf. Pfeffermann, 2013). Besides these formal procedures, also graphical devices may be used in model selection. An excellent example can be found in the paper due to Verret, Hidiroglou, and Rao (2015), where the level one residuals are plotted against variables omitted from the model. If such a plot indicates a linear relationship between the residuals and the omitted variable, the prediction can be improved by including the latter in the model.

3.5.3 Applications

After introducing the tools for diagnosis and selection, we apply them to two commonly used data sets. The most widely analysed data set in small area estimation for unit level models is the area under crop in certain counties of Iowa introduced by Battese et al. (1988). The authors estimated the areas under corn and soy-beans for twelve counties, using satellite data as auxiliary information. For detailed analyses of the modelling, see Battese et al. (1988), Jiang and Lahiri (2006b, Section 6) and Rao (2003, Section 7.2.1). In the following analysis, we estimate the variance components using REML, while in the original article a fitting-of-constants approach was adopted. For the estimation of the area under corn, we begin with the same model as Battese et al. (1988), i.e. we use a random intercept model with the satellite data on areas under corn and soy-beans as predictors:

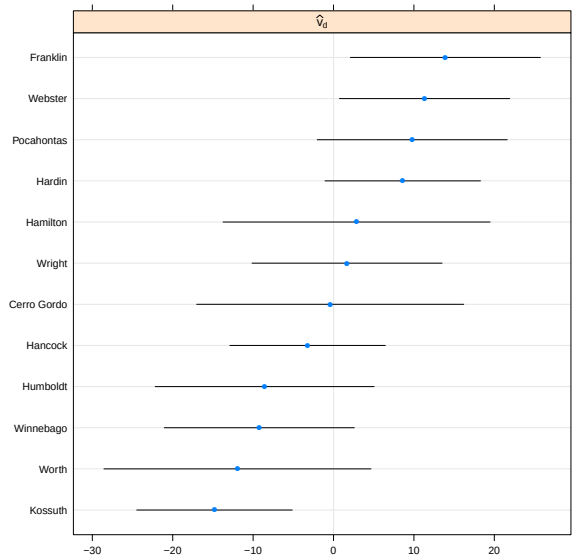

Figure 3.1: $\widehat{v}_d$ for the corn data

$$y_{dj} = \beta_0 + \beta_1 x_{dj1} + \beta_2 x_{dj2} + v_d + \varepsilon_{dj}, \quad d = 1, \ldots, n_d,\ d = 1, \ldots, D,$$

where $x_{dji},\ i = 1, 2$ denotes the number of pixels classified as corn and soy-beans from the satellite images. For the v_d, two extreme options are to model them as fixed effects or to require that $v_d = 0$ for all d, which amounts to not modelling any unobserved heterogeneity at all. A graphical examination of the relationship led Battese et al. (1988) to reason that a linear model is appropriate. As pointed out by Jiang and Lahiri (2006b), the sample sizes in the areas are insufficiently small to treat the v_d as fixed effects, which is why Battese et al. (1988) modelled them as interchangeable realisations of a random variable. Hence, rather than considering a random intercept model to be an accurate description of the truth, it was taken as a reasonable modelling compromise. To assess whether a random effects model is reasonable, the EBLUPs $\widehat{v}_d$ can be plotted along with prediction intervals obtained from the prediction variance, i.e. the g_{1d}-term present in the MSE (3.2.12). This is depicted in Figure 3.1, where the blue points indicate the EBLUP for the respective area, which is surrounded by the 95% prediction interval. It can be seen that in the counties of Kossuth, Franklin and Webster, the prediction intervals do not contain 0, but only closely so for Franklin and Webster. Moreover, we do not see evidence of an outlier in the random effects. Altogether, the i.i.d. assumption on the v_d does not seem implausible in this model. Furthermore, a Shapiro-Wilk test of normality of the $\widehat{v}_d$ yields a test statistic of $W = 0.9462$ corresponding to a p-value of 0.5821. Hence, normality on behalf of the random effects cannot be rejected. Besides, QQ plots and associated Shapiro-Wilk tests of normality for the transformed residuals do not indicate problems with the normality assumption on the ε_{dj} either. The same holds for the plot of the standardised residuals against the predicted responses $\widehat{y}_{dj}$, which is shown in Figure 3.2. Hence, the diagnostics do not highlight severe problems with the assumption of a random intercept model. Nonetheless, we might still be interested whether other modelling options could be preferred. To compare the random intercept model against a complete pooling model assuming $\sigma_v^2 = 0$, we perform an exact RLRT due to Crainiceanu and Ruppert (2004). This procedure yields a test statistic of 7.7 with p-values around 0.002, which indicates

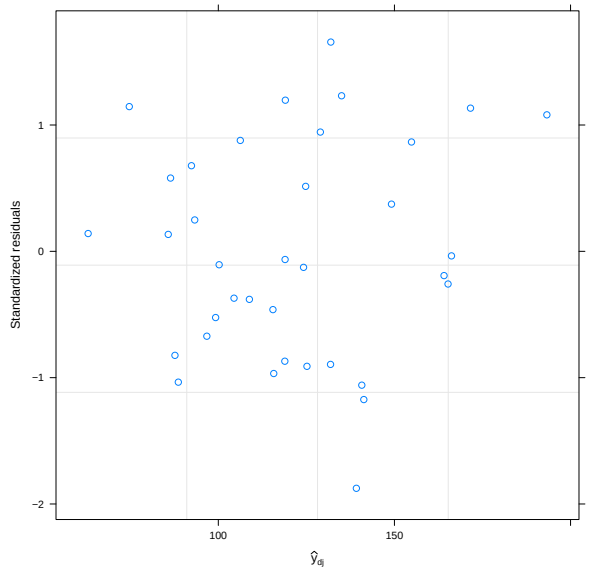

Figure 3.2: Standardised residuals against fitted values for the corn data

that we should not disregard the random effect. Moreover, we can compare the predictive accuracy of the different modelling options regarding the domain effects. To do so, we use the marginal and conditional AIC given in equations (3.5.1) and (3.5.2). The marginal AIC for the complete pooling and fixed effects case amounts to 309.5797 and 294.7113 respectively, implying that it is much better to incorporate the areas as fixed effects rather than ignoring them altogether. The value of the cAIC for the random intercept model is 295.5543. Hence, interestingly, according to the (c)AIC, there is a slight advantage of modelling the domain effects as fixed rather than as random. However, following Vaida and Blanchard (2005, p. 360), we may argue that with a small difference in (c)AIC between two models, this criterion should not be used to select between two models. In this case, using fixed effects for some areas with only one observation does not seem to be a good idea, anyway. Repeating the above analysis for the model with the area under soy-beans as response variable shows that models without the satellite information on the pixels of corn as a covariate yield a higher predictive accuracy. These results are not fully surprising, since a scatter plot of the reported area under soy-beans versus the pixels of corn does not support a linear relationship between these two quantities.

Another interesting application is the true average expenditure on fresh whole milk, which is estimated by the US bureau of labour statistics for 43 areas using an area level model. The data set is provided in the R-package *sae* by Molina and Marhuenda (2013), and details are available in the papers of Arora and Lahiri (1997) and You and Chapman (2006). It should be noted that in both these papers, a fully Bayesian approach has been adopted, whereas we will use the frequentist framework. Following You and Chapman (2006), and the procedure in the example of the R-package, the four major areas will be used as auxiliary information in the model. Thus the model reads

$$\widehat{\mu}_d^{Dir} = \beta_0 + \beta_1 x_{d1} + \beta_2 x_{d2} + \beta_3 x_{d3} + v_d + \varepsilon_d, \quad d = 1, \ldots, 43, \tag{3.5.3}$$

where $x_{di},\ i = 1, 2, 3$ indicates the dummies for major areas 2 to 4 and $\widehat{\mu}_d^{Dir}$ denotes the

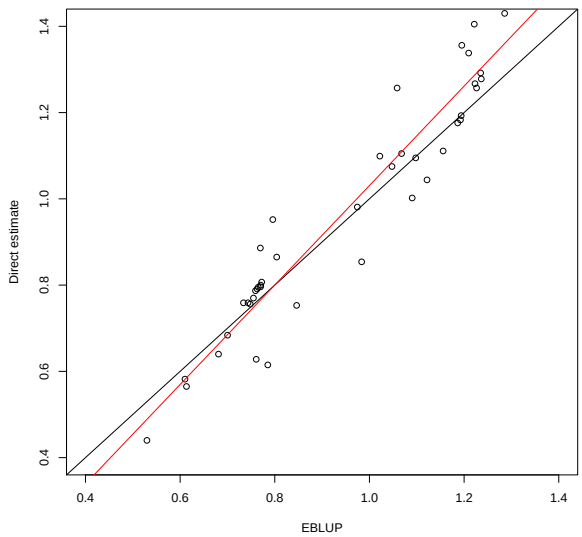

Figure 3.3: Design-based versus model-based estimates for the milk data

direct estimate for area d. First we look at the graphical bias diagnostic due to Brown et al. (2001), which is depicted in Figure 3.3. The black line indicates the 45 degree line and the red line the estimated regression line. The graph does not indicate a clear signal for or against the use of the EBLUP in this application. Rather the shrinkage property of the EBLUP is evident, especially for the areas with relative large or small values of the direct estimator. Besides an F-test can performed to asses the restriction that in the regression of the direct estimates against the EBLUP, the intercept is 0 and the slope 1. The F-statistic takes the value of 4.9925, which corresponds to a p-value of 0.01147 using the F(2,41) distribution. Hence, it is unlikely that these restrictions hold. Since the $\boldsymbol{\beta}$-vector for the Fay-Herriot model is obtained from a weighted least squares approach, standard influence diagnostics are typically desired as well. Figure 3.4 shows in the upper graph the studentised residuals from a regression with the d-th area removed. In the middle part Cook's distance is depicted, whereas the graph at the bottom indicates the estimated random effects variance, when area d is deleted. Interestingly, all three graphs indicate that areas 4 and 11 are influential, i.e. the estimated model parameters change drastically if either of these areas is removed from the model fitting. Moreover, the standardised residuals $(\widehat{\mu}_d^{Dir} - \widehat{\mu}_d^{FH})/\sqrt{\psi_d}$ can be examined. A Shapiro-Wilk test yielded a test statistic of $W = 0.9611$ corresponding to a p-value of 0.1522, i.e. normality cannot be rejected for the level one residuals. A plot of the standardised residuals against the EBLUPs under model (3.5.3) displayed in Figure 3.5, reveals a trend. These findings indicate some potential problems with the model at hand. Since further auxiliary data are not available, the degree to which model selection can be performed is limited. A comparison of model (3.5.3) against a model without the major areas as auxiliary information using the cAIC due to Han (2013), indicates a clear advantage for model (3.5.3). While model (3.5.3) yields a value for the cAIC of 66.0457, the reduced model without covariates yields a cAIC-value of 75.48596.

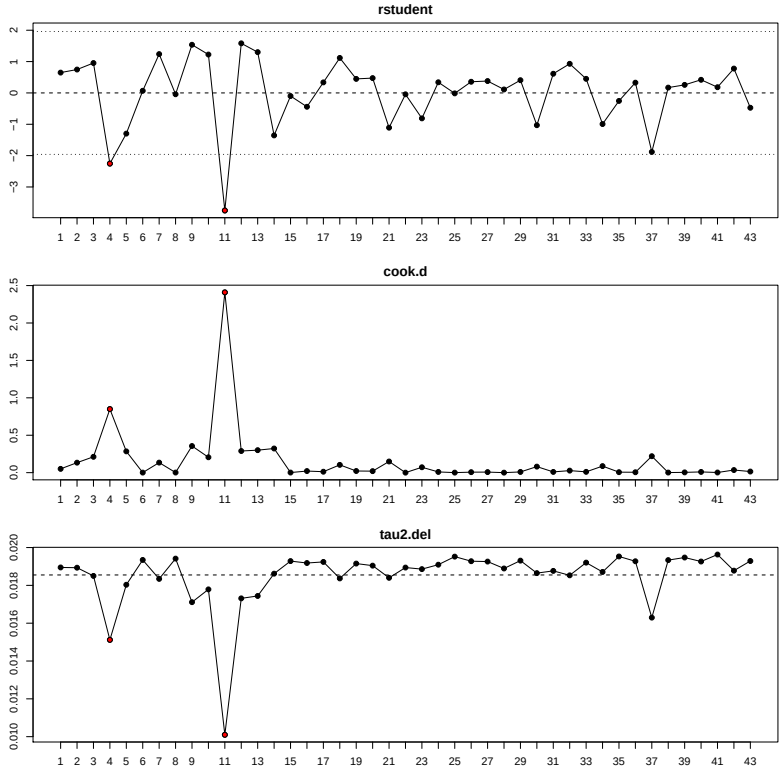

Figure 3.4: Influence diagnostics for the milk data

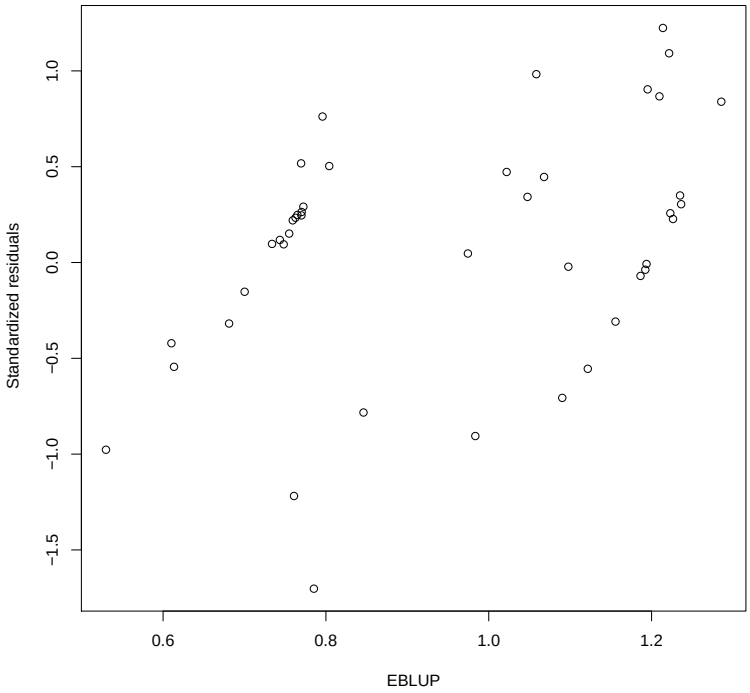

Figure 3.5: Standardised residuals against FH-estimates for the milk data

3.6 Summary and discussion

This chapter introduced model-based techniques for small area estimation. After presenting the general linear mixed model, the two most commonly used small area models due to Fay and Herriot (1979) and Battese et al. (1988) were introduced. A drawback of the latter approach is that it ignores the survey design altogether. Hence, it is not design-consistent for general sampling designs (cf. Rao, 2003, Section 7.2.1) and it may be subject to a sample selection bias. In subsequent chapters, estimation procedures which are either design-consistent or avoid a selection bias will be presented. A slightly more subtle issue concerns the practical relevance of the model-unbiasedness of the EBLUPs. As pointed out by Pfeffermann (2013, Remark 1), predictors (3.2.8) and (3.2.21) are biased when conditioning on v_d. In applications of small area estimation, however, the aim is to estimate means or other quantities of a particular population, which are conditional on the realised random effects (Pfeffermann, 2014).

In Sections 3.3 and 3.4 generalisations of a linear mixed model were considered by introducing non-linear transformations of the response, such that the assumed random intercept model holds for the transformed data. A prerequisite for the successful implementation of any of those techniques is the validity of the assumed model which was discussed next. So far the presentation of model-based procedures focused on situations where the response variable is continuous, but in many applications the response may be binary, e.g. whether or not a household is poor. In these cases y_{dj} will be Bernoulli-distributed, where (McCullagh & Nelder, 1989, chapter 4)

$$Pr(y_{dj} = 1) = p_{dj} \text{ and } Pr(y_{dj} = 0) = 1 - p_{dj}. \tag{3.6.1}$$

Since the probabilities p_{dj} fulfil $0 \leq p_{dj} \leq 1$, a link function $g(\cdot)$ is applied, such that $g(p_{dj})$ takes values on the real line and thus can be linearly related to a set of covariates $\mathbf{x}_{dj}$. The most common link functions are the logit function

$$g_1(p_{dj}) = \log\left(\frac{p_{dj}}{1 - p_{dj}}\right), \tag{3.6.2}$$

and the probit function

$$g_2(p_{dj}) = \Phi^{-1}(p_{dj}), \tag{3.6.3}$$

where $\Phi^{-1}(\cdot)$ denotes the inverse of the distribution function of the normal distribution (McCullagh & Nelder, 1989, p. 108). The first attempts to apply logistic mixed models to small domain estimation problems were conducted within a Bayesian framework. A hierarchical Bayes (HB) approach using the Gibbs sampler and accounting for spatial correlation was proposed by Ghosh, Natarajan, Stroud, and Carlin (1998). A detailed account on various ways of modelling proportions by HB area level models is given by Liu (2009). Jiang and Lahiri (2001) note that the HB approach albeit being very flexible is highly computationally demanding, which is why they introduced an empirical best prediction approach for small area estimation with binary data. They consider the following model

$$
\begin{aligned}
y_{dj}|p_{dj} &\overset{ind}{\sim} Bern(p_{dj}) \\
logit(p_{dj})|v_d &= \mathbf{x}_{dj}^T\boldsymbol{\beta} + v_d \\
v_d &\overset{i.i.d.}{\sim} N(0, \sigma_v^2),
\end{aligned}
\tag{3.6.4}
$$

where $Bern(\cdot)$ denotes the Bernoulli distribution. The MSE-optimal predictor for the non-sampled units under this model emerges as the ratio of two one-dimensional integrals, which can be evaluated using numerical integration (Pfeffermann & Correa, 2012). A naive plug-in predictor for the mean μ_d may be constructed from (3.6.4) as

$$
\begin{aligned}
\widehat{\mu}_d^{naive} &= \frac{1}{N_d}\left[\sum_{j\in S_d} y_{dj} + \sum_{j\notin S_d} \widehat{p}_{dj}^{naive}\right] \qquad \text{where} \\
\widehat{p}_{dj}^{naive} &= \frac{\exp(\mathbf{x}_{dj}^T\widehat{\boldsymbol{\beta}} + \widehat{v}_d^{naive})}{1+\exp(\mathbf{x}_{dj}^T\widehat{\boldsymbol{\beta}} + \widehat{v}_d^{naive})},
\end{aligned}
\tag{3.6.5}
$$

where $\widehat{v}_d^{naive}$ is obtained from the model (3.6.4). Since the MSE-optimal predictor under (3.6.4) does not have a closed-form expression, however, simpler expressions such as (3.6.5) are popular in practice. González-Manteiga, Lombardía, Molina, Morales, and Santamaría (2007) coined the term "logistic mixed model predictor" for a similar plug-in predictor where additionally a unit-specific error term is modelled for the logit coefficients of p_{dj} which accounts for overdispersion. They also covered MSE estimation for their procedure by means of a Taylor linearisation and a wild bootstrap.

Moreover, model-based small area procedures have also been proposed for general categorical y_{dj}. Examples in this regard can be found in López-Vizcaíno, Lombardía, and Morales (2013); Molina, Saei, and Lombardia (2007); Saei and Taylor (2012).

The strategies for MSE estimation under the linear mixed model discussed in Sections 3.1 and 3.2, mainly focused on approximations based on Taylor linearisation. A drawback of these procedures is that they critically depend on the normality assumption, which may not be valid in many applications. Moreover, the implementation of a Taylor series approximation can be very cumbersome for more complicated models, such as non-linear mixed models. For those cases resampling-based approaches may avoid the necessity of deriving approximations (cf. S. N. Lahiri et al. (2007)). These approaches are now feasible due to the advances in computer power over the last decade and promoted research in fields such as MSE estimation of EBPs and the construction of valid prediction intervals. The most commonly used resampling procedures in this regard are bootstrap and jackknife techniques. An excellent overview about their use in small area estimation is given in Rao (2007). Following Berg and Chandra (2014, p. 161), the MSE of an EBP can be written as the sum of the prediction variance of the BP and the squared difference between the EBP and the BP. This implies (cf. Rao, 2007)

$$
\begin{aligned}
\mathrm{MSE}(\widehat{\mu}_d^{EBP}) &= \mathrm{E}(\widehat{\mu}_d^{EBP} - \mu_d)^2 \\
&= \mathrm{E}(\widetilde{\mu}_d^{BP} - \mu_d)^2 + \mathrm{E}(\widehat{\mu}_d^{EBP} - \widetilde{\mu}_d^{BP})^2 \\
&= M_{1d}(\boldsymbol{\zeta}) + M_{2d}(\boldsymbol{\zeta}),
\end{aligned}
\tag{3.6.6}
$$

where $M_{1d}(\boldsymbol{\zeta})$ is the leading term and $M_{2d}(\boldsymbol{\zeta})$ captures the uncertainty due to the estimation of the model parameters (Berg & Chandra, 2014, p. 161). Note that the MSE decomposition was already used for the linear mixed model, see (3.1.22), but the EBP methodology is more general. It includes among others the EBP under a logistic mixed model, the predictor proposed by Berg and Chandra (2014) under a lognormal mixed model and also the EBLUPs discussed in Section 3.2, provided distributional assumptions with respect to the error terms are made. Note that simply replacing $M_{1d}(\boldsymbol{\zeta})$ by $M_{1d}(\widehat{\boldsymbol{\zeta}})$ will lead to a bias, which would mask the contribution of the $M_{2d}(\cdot)$-term. To estimate the components $M_{1d}(\boldsymbol{\zeta})$ and $M_{2d}(\boldsymbol{\zeta})$, Jiang et al. (2002) (JLW, henceforth) proposed a jackknife, where a bias correction for the leading term is included, which leads to a nearly unbiased MSE estimator. The required steps for computing MSE estimates using the JLW-methodology are detailed in Rao (2003, Section 9.2.2). Alternatively, parametric bootstrap procedures can be used to obtain bias-corrected estimates of $M_{1d}(\widehat{\boldsymbol{\zeta}})$ and $M_{2d}(\widehat{\boldsymbol{\zeta}})$ (for further details, see Pfeffermann & Glickman, 2004 and Rao, 2007, Section 4). The JLW approach assumes that the model parameters are estimated using M-estimators, which is generally the case, as this class comprises ML / REML and ANOVA approaches. The authors conduct several simulation studies which indicate a strong performance of their approach relative to the MSE estimator based on Taylor series arguments. It has been pointed out by Rao (2007) that the JLW-jackknife requires an analytical expression for the leading term of the MSE estimator. Provided that this is the case, the JLW approach is a powerful strategy. However, due to the bias correction for the leading term by the jackknife, negative MSE estimates could occur (cf. Rao, 2007 and the references therein for strategies to avoid this issue). Lohr and Rao (2009) (LR) considered a modification of the JLW approach, which is both area-specific and also computationally simpler. The former implies that the LR-jackknife tracks the conditional MSE, which also yields an unbiased estimation of the unconditional MSE as pointed out by Lohr and Rao (2009). The use of the conditional MSE was suggested by Booth and Hobert (1998), who argued that inference on the random effects should be based on their conditional distribution given the data. As pointed out by Booth and Hobert (1998, p. 265), this requirement is not met by the unconditional MSE unless the conditional variance is constant, which is the case for the normal mixed models covered in Section 3.2. Furthermore, Zhang (2007) investigated the issue of interval estimation for a finite population. He showed that the area-specific coverage of the BP under model (3.2.17) degenerates in the important case that $\sigma_v^2/\psi_d \approx 0$. As an alternative, he proposes to consider the simultaneous coverage, which averages over the area-specific coverages.

Another important issue is with regards to the coherence of estimates at different levels of aggregation. As pointed out by Wang, Fuller, and Qu (2008), practitioners may require that the weighted sum of model-based small area estimates reduces to the national statistic estimated by a design-based or model-assisted procedure. This implies

$$\sum_{d=1}^{D} N_d \widehat{\mu}_d = \widehat{\tau}, \tag{3.6.7}$$

where the $\widehat{\mu}_d$ are model-based domain estimates and $\widehat{\tau}$ is a design-consistent estimate of the national total. An estimator satisfying condition (3.6.7) is said to be benchmarked or self-calibrated against the national total (cf. Wang et al., 2008 and references therein). Note that neither the EBLUP under the unit level model (3.2.7) nor the EBLUP under the

area level model (3.2.21) satisfy (3.6.7) without further adjustments. You and Rao (2002a) proposed a pseudo-EBLUP based on the unit level model which fulfils the benchmarking property automatically. This estimator will be discussed in detail in Section 4.2.1. Wang et al. (2008) applied the idea due to You and Rao (2002a) to the area level model (3.2.17), yielding a self-calibrated estimator of the area means. You, Rao, and Hidiroglou (2013), who derived a corresponding MSE estimator, noted that imposing the benchmarking requirement will lead to slightly higher MSEs compared to the respective EBLUP. Hence, the consequences of imposing restriction (3.6.7) are ambiguous. On the one hand, it leads to a modification of the MSE optimal predictor, causing an increase in the MSE if the assumed model holds. On the other hand, the self-calibration property is very desirable from the viewpoint of national statistical institutes. They produce statistics at different levels of aggregation and have to ensure that these published figures are coherent with each other. Thus, a violation of property (3.6.7) may be deemed unacceptable by data producers, as it could lead to a distrust in their statistics. A very interesting proposal in this regard is due to Münnich, Sachs, and Wagner (2012), who invoke calibration constraints on different levels of aggregation. Since it may not be possible to satisfy all benchmarks at once, the authors allow the values to differ, but penalise those deviations. Besides achieving coherence, their approach also yields insights with regards to the price the user has to pay in terms of perturbing the model-based estimates to satisfy a given constraint.

Chapter 4

Accounting for the sampling design in model-based small area estimation

It has been argued in the previous chapter on model-based estimation that once a model has been validated to the sample data, valid predictions for the quantities of interest can be obtained provided the sample is non-informative. Thus, this chapter focusses on strategies dealing with informative sampling designs. We introduce this problem and review some tests to detect informativeness in Section 4.1. Section 4.2 is devoted to methods which account for the sampling design. These include approaches imposing design consistency for model-based small area procedures (Prasad & Rao, 1999; You & Rao, 2002a) and methods alleviating the bias due to informative sampling (Pfeffermann & Sverchkov, 2007; Verret et al., 2015). These strategies were developed for cases where domain estimates were desired for the dependent variable in a unit level nested error regression model. Recently, model-based procedures for small area estimation when the dependent variable is log-transformed have been developed (Berg & Chandra, 2014), as discussed in Section 3.4. If the sampling design is non-ignorable, adjustments to the EBP of Berg and Chandra may be called for, e.g. incorporating survey weights or augmenting the auxiliary information by the a function of the selection probabilities. We propose different approaches to extend the EBP in Section 4.3. We elaborate on the findings of a simulation study on the performance of these adjustments to the EBP of Berg and Chandra in Section 4.4. The main insights of this chapter are summarised in Section 4.5.

4.1 The issue of informative sampling

4.1.1 Background on informative sampling

It is commonly agreed that the two main purposes why models are fitted to survey data are analytic and descriptive inference (cf. Chambers & Skinner, 2003a; Pfeffermann, 2011; Pfeffermann & Sverchkov, 2009; Rao, Hidiroglou, Yung, & Kovacevic, 2010; Skinner, 1994). Whereas in the former approach the main interest lies on the model parameters, the latter approach focusses on functions of finite population quantities, such as the prediction of means in small area estimation (cf. Pfeffermann, 2011, p. 115). Intuitively, the issue of informative sampling arises when a model can be validated for the sample at hand, but

the parameters are different from the ones holding for the population (cf. Pfeffermann & Sverchkov, 2009, p. 455). Since the ultimate aim is to draw inferences about population quantities, such as the regression model in case of analytic inference or finite population totals in descriptive inference, this may yield misleading conclusions. As pointed out by Pfeffermann and Sverchkov (2009, p. 456), these biases may arise due to the sampling process itself, the response behaviour or both. Hence, even if the sampling design is completely non-informative, an informative response mechanism may result in erroneous inferences when the sampling model is applied (see the discussion in Valliant, Dorfman, & Royall, 2000, Section 2.6.2). From now on, we shall assume that we are in the lucky situation of full response. Following Pfeffermann and Sverchkov (2009, p. 456 ff.), we may assume a single level regression model, such that the marginal conditional density function of y_k given the covariates $\mathbf{x}_k$ and the sample membership $I_k = 1$ can be expressed as:

$$f_S(y_k|\mathbf{x}_k) = f_U(y_k|\mathbf{x}_k, I_k = 1) = \frac{\Pr(I_k = 1|y_k, \mathbf{x}_k) f_U(y_k|\mathbf{x}_k)}{\Pr(I_k = 1|\mathbf{x}_k)}. \tag{4.1.1}$$

In the above equation $f_S(\cdot)$ denotes the sample density function and $f_U(\cdot)$ the corresponding population density function. From (4.1.1) it is evident that the sample density function can be used to draw inferences on the population density function without further adjustments if the following condition holds:

$$\Pr(I_k = 1|y_k, \mathbf{x}_k) = \Pr(I_k = 1|\mathbf{x}_k) \forall y_k. \tag{4.1.2}$$

This condition implies that conditional on the covariates the response and the selection are independent. As illustrated by Valliant et al. (2000, Section 2.6.1), cut-off sampling based on the values y_k violates condition (4.1.2) and thus yields informative samples. Cut-off sampling can also be considered an extreme case of endogenous stratification, where one stratum contains all elements with $y_k > C$ and all other elements belong to the second stratum. In this case, the inclusion probabilities satisfy $\pi_k = \mathbb{I}(y_k > C)$, where C refers to the known cut-off threshold (cf. Chambers, 2003). Incidentally, cut-off sampling does not yield a probability sample as the requirement $\pi_k > 0$ is not met for all units (cf. Section 2.1 of this thesis and the references therein).

There are also many sampling schemes which satisfy condition (4.1.2), such as simple random sampling or when the inclusion probabilities are a function of the covariates, i.e. $\Pr(I_k = 1|y_k, \mathbf{x}_k) = g(\mathbf{x}_k)$, as is the case in sampling with probability proportional to size if the size variable is included among the covariates. Note that there are cases where the use of the sample distribution may be valid for the estimation of some model parameters, even if (4.1.2) does not hold. An example is given by Pfeffermann and Sverchkov (2009, Remark 1.2). In small area estimation, however, the aim is to predict area quantities such as the mean. Hence, a sampling scheme which yields unbiased estimates of all $\boldsymbol{\beta}$ components but the intercept will produce a biased estimate of the target quantity. This is important since it has been pointed out by Fuller (2009, p. 352) that the intercept term may suffer from the highest bias in practice.

To illustrate the bias due to an informative sampling mechanism, we consider the estimation of a finite population total, which is constructed under the model

$$\begin{aligned} y_k &= \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k \\ \varepsilon_k &\overset{i.i.d}{\sim} N(0, \sigma^2). \end{aligned} \tag{4.1.3}$$

Owing to the independence of the error term the BLUP of the total τ follows as (Valliant et al., 2000, Corollary 2.2.1)

$$\widehat{\tau}^{BLUP} = \sum_{k \in S} y_k + \sum_{k \in U \setminus S} \mathbf{x}_k^T \widehat{\boldsymbol{\beta}}. \tag{4.1.4}$$

An implicit assumption regarding the BLUP is that the sampling design is non-informative. Suppose, however, that we have two strata (A and B) and the stratum membership is related to the outcome variable in the following manner:

$$\Pr(k \in A) = \begin{cases} 0.75, & \text{if } y_k > y_{[0.5]} \\ 0.25, & \text{otherwise,} \end{cases} \tag{4.1.5}$$

where $y_{[0.5]}$ denotes the median of the variable of interest. Furthermore, the sampling fraction in stratum A is four-times the sampling fraction in stratum B. Since the outcome variable depends on the error term, the resulting sampling scheme has traits of an endogenous stratification and is clearly informative. In this setting the stratum membership can be viewed as a kind of prior information whether a unit is large or not. To show what can go wrong, we construct finite populations of size $N = 10000$ according to model 4.1.3 where the covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ are drawn once and then held fixed. They are drawn from $\mathbf{x}_1 \overset{i.i.d}{\sim} N(1,1)$ and $\mathbf{x}_2 \overset{i.i.d}{\sim} U[0,1]$. The model parameters are set to $\boldsymbol{\beta} = (1, 2, -1)^T$ and $\sigma^2 = 2$. The setting is a finite population model-based simulation with $R = 10000$ replications and a sample size of $n = 500$. For comparison, we also draw samples with UPS with πps, using $\mathbf{x}_2$ as the size variable. Note that this procedure yields a high variation of design weights. Using properties of the distribution of the maximum and minimum of independently and identically distributed random variables, we can infer the ratio of the median of the maximum to the median of the minimum of the design weights as $0.5^{1/N}/(1 - 0.5^{1/N}) = 14426.45$. Nonetheless, since the inclusion probabilities are a function of $\mathbf{x}_2$, which is included in the model, we do not expect a bias under the UPS design. The results of this small simulation study are given in Table 4.1.

Table 4.1: Relative biases for an informative and non-informative sampling design

	$\widehat{\tau}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
UPS	−0.0004	−0.0061	0.0007	−0.0078
StrRS	0.1381	0.3921	−0.0296	−0.0253

It can be seen that the UPS design yields very good estimates of the total, while the StrRS scheme overestimates the true total on average by almost 14 per cent. These results highlight that a non-informative sampling process does not infer with the model, provided that it is correctly specified, even if there is a large variation of the weights. The last 3 columns indicate that the Monte-Carlo relative biases for the model parameters (measured against the true vector of regression coefficients) are very close to zero as well, i.e. under the UPS design, valid point estimates are achieved because the model is correctly estimated. For the StrRS scheme, the bias in the point estimate can be largely ascribed to the severe overestimation of the intercept term, confirming Fuller's insight.

It has been noted above that equations (4.1.1) and (4.1.2) are derived using a single level model. For small area estimation under the random intercept model (3.2.1), the two-stage sampling design summarised in Algorithm 2 is frequently assumed (cf. Pfeffermann & Sverchkov, 2007; Verret et al., 2015).

Algorithm 2 Two-stage design for model-based small area estimation

1. A subset $m \subseteq D$ of the domains are sampled at the first stage with inclusion probabilities $\pi_d,\ d = 1, \ldots, D$;
2. within the sampled domains a sample of (expected) size n_d is taken using second-stage inclusion probabilities $\pi_{j|d},\ j = 1, \ldots, N_d$.

It should be noted that although the domains are incorporated in the sampling design, this is not a good setup for design-based and model-assisted estimation techniques, since there will be non-sampled domains, unless $m = D$. If this holds, the domains are design strata and we are back to the planned domain case discussed in Chapter 2. Instead, this two-stage sample design may be chosen due to practical reasons, such as budget constraints. With this two-stage design, we need to distinguish between informativeness of the first stage and second stage. Following Pfeffermann and Sverchkov (2007, p. 1428), the conditions under which the sampling design is ignorable at level two and one are given by

$$\begin{aligned} \Pr(I_d = 1|v_d) &= \Pr(I_d = 1) \\ \Pr(I_{dj} = 1|y_{dj}, \mathbf{x}_{dj}, v_d, I_d = 1) &= \Pr(I_{dj} = 1|\mathbf{x}_{dj}, v_d, I_d = 1). \end{aligned} \tag{4.1.6}$$

Hence, the condition on the second level requires the sampling of the domains to be independent of the random effects. The condition for level one ignorability is similar to (4.1.2) and states that after accounting for the covariates $\mathbf{x}_{dj}$ the sample membership should be independent of the response.

An excellent account on the implications of a non-informative sampling design for small area estimation is given in Rao (2003, Section 5.3). The use of the sampling model to draw inferences on $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \sigma_v^2, \sigma_\varepsilon^2)^T$ is justified provided that the inclusion probabilities do not depend on the response vector in any of the domains. This is important, since the vector $\widehat{\boldsymbol{\xi}}$ is used to obtain predictions for the non-sampled elements. Hence, a bias in $\widehat{\boldsymbol{\xi}}$ may yield a bias in the small area quantity to be estimated. Moreover, Rao (2003, Section 5.3) points out that if the selection probabilities depend on a vector $\mathbf{z}_d$, which is not included in the covariates, this will lead to an informative sampling design as long as $\mathbf{y}_d$ and $\mathbf{z}_d$ are not independent conditional on the model covariates.

4.1.2 Testing for informativeness

Since the informativeness of a sampling design may invalidate the conclusions altogether, the detection of non-ignorable sampling designs is crucial for the successful implementation of model-based estimation techniques. As a result, many testing procedures have been proposed in the literature. With the purpose of analytic inference in mind, Skinner (1994) proposes to regress the variable of interest against the model covariates, the survey weights

and, possibly, interaction terms. He would then test whether these additional regression coefficients related to the survey design are jointly zero. Alternatively, this procedure can be derived as well from a comparison of weighted and unweighted estimates, see Fuller (2009, Section 6.3.1) and the references therein for details. Moreover, it has been argued that this test can be a good diagnostic to detect important design variables (cf. Pfeffermann, 1993; Skinner, 1994 for a discussion and references). In a similar vein, a Hausman-type test was proposed by Pfeffermann (1993) to test the ignorability of a design. Suppose that $\widehat{\boldsymbol{\beta}}_0$ is the efficient maximum-likelihood estimator under some model whereas $\widehat{\boldsymbol{\beta}}_w$ is a design-consistent estimator. Now since $\widehat{\boldsymbol{\beta}}_0$ is asymptotically normally distributed and assuming the same holds for $\widehat{\boldsymbol{\beta}}_w$ under the null hypothesis that the design can be ignored the test statistic H given by

$$H = \left(\widehat{\boldsymbol{\beta}}_w - \widehat{\boldsymbol{\beta}}_0\right)^T \left[\widehat{V}\left(\widehat{\boldsymbol{\beta}}_w\right) - \widehat{V}\left(\widehat{\boldsymbol{\beta}}_0\right)\right]^{-1} \left(\widehat{\boldsymbol{\beta}}_w - \widehat{\boldsymbol{\beta}}_0\right) \tag{4.1.7}$$

follows a χ^2_p distribution, where p denotes the number of covariates included in the model (Pfeffermann, 1993). Asparouhov (2006) considered test statistics which are similar to (4.1.7) but aimed at a single response variable rather than a vector of regression parameters. The first statistic they evaluated is

$$I_1 = \frac{\widehat{\mu}_w - \widehat{\mu}_0}{\sqrt{\widehat{\sigma}^2_w - \widehat{\sigma}^2_0}}, \tag{4.1.8}$$

where $\widehat{\mu}_w(\widehat{\sigma}^2_w)$ and $\widehat{\mu}_0(\widehat{\sigma}^2_0)$ are the estimated weighted and unweighted means of the variable of interest (estimated variances of the mean of this variable), respectively. A higher absolute value of I_1 indicates informative sampling. Chambers, Dorfman, and Sverchkov (2003) proposed a test similar to (4.1.8), which is applicable for non-parametric regression models, where the difference of the variances in the denominator is estimated using a jackknife. However, as noted by Asparouhov (2006), as the variances will decrease with an increasing sample size, I_1 will likely indicate informativeness for large samples, despite small differences between the point estimates. Also Chambers et al. (2003) report unsatisfactory results using their approach. Furthermore, the test statistic (4.1.8) is not defined whenever $\widehat{\sigma}^2_w \leq \widehat{\sigma}^2_0$. To overcome these shortcomings, Asparouhov (2006) investigated the use of the statistic

$$I_2 = \frac{|\widehat{\mu}_w - \widehat{\mu}_0|}{\sqrt{\widehat{v}^2_0}}, \tag{4.1.9}$$

where $\widehat{v}^2_0$ denotes the estimated variance of the response variable. I_2 is independent of the sample size but may be affected by the scaling of the weights. To get rid of this dependence, Asparouhov (2006) recommends using

$$I_3 = \frac{|\mu - \widehat{\mu}_0|}{\sqrt{\widehat{v}^2_0}}, \tag{4.1.10}$$

with μ as the true mean of the response variable. Note that none of these statistics, I_1, I_2 and I_3 allows to distinguish between informativeness on the different levels. Moreover, Chambers et al. (2003) propose to perform correlation tests of the weighted residuals against the polynomial of the single auxiliary variable in non-parametric regression models. Alternatively, as a sample design is ignorable if condition (4.1.6) holds, a straightforward idea is to use these conditions in a test procedure. With this notion in mind, Pfeffermann and Sverchkov (2007) propose tests specific for the two-level model (3.2.1) that allow to check whether the sampling design on either level is informative. Pfeffermann and

Sverchkov (2007) derive that a test of informativeness of the sampling of areas can be implemented by regressing the area weights w_d against the predicted random effects $\widehat{v}_d$. Provided that the error terms of this regression are i.i.d. normally distributed, the t-statistic of the slope follows a t-distribution with $m-2$ degrees of freedom. To test whether the within-areas sample designs are informative, the regression coefficient on y_{dj} from regressing the conditional level one survey weights $w_{j|d}$ on $\mathbf{x}_{dj}$ and y_{dj} could be considered. But, as noted by Pfeffermann and Sverchkov (2007), this procedure may suffer from small sample sizes within areas and therefore have low power. Instead, they recommend using the maximum value of the test statistic over all sampled areas, assuming a common within-areas design. This statistic can then be compared to the critical value obtained from the distribution function of the maximum.

4.2 Accounting for the sampling design under a linear unit level model

4.2.1 Weighted estimation of the model parameters

There are various ways of incorporating the sample design in the modelling of survey data. A general discussion of this topic can be found in Pfeffermann (1993). One approach considers pseudo-ML methods, where design weights enter into the likelihood function. For mixed models, which are commonly applied in small area estimation, Pfeffermann, Skinner, Holmes, Goldstein, and Rasbash (1998) developed an extension of the iteratively generalised least squares algorithm which accounts for survey weights. They assume a two-stage sampling design as described in Algorithm 2, where on the first stage areas are sampled and then within the sampled areas a sample of the level one units is drawn. In the case of planned domains as detailed in Chapter 2, the areas are considered as strata, which are sampled with certainty. As a consequence the domain-level weights simplify to $w_d \equiv 1\ \forall d$ and therefore, $w_{j|d} = w_{dj}$. Hence, we can directly focus on survey weights w_{dj} for unit j in domain d, which are assumed to be the inverse of the inclusion probabilities, i.e. $w_{dj} = \pi_{dj}^{-1}$. The algorithm due to Pfeffermann et al. (1998) uses the block-diagonal covariance structure and therefore splits the problem of estimating $\boldsymbol{\xi}$ into a summation within domains and then a sum over these domain sums in a second step. Details about the implementation can be found in Pfeffermann et al. (1998, Section 4). Furthermore, the authors show that their procedure yields design-consistent estimates $\widehat{\boldsymbol{\beta}}_w$ and $\widehat{\boldsymbol{\delta}}_w = (\sigma^2_{v,w}, \sigma^2_{\varepsilon,w})^T$. Besides, as noted by Pfeffermann et al. (1998), the asymptotic model-unbiasedness of $\widehat{\boldsymbol{\xi}}_w = (\widehat{\boldsymbol{\beta}}_w^T, \widehat{\boldsymbol{\delta}}_w^T)^T$ is not affected by scaling the weights. Hence, the conditional level one weights may be scaled as $w^*_{j|d} = \lambda_d w_{j|d}$ as to avoid small sample biases. Different choices for the scaling factor λ_d may be considered. As an example Pfeffermann et al. (1998) consider

$$\lambda_d = (\overline{w}_{j|d})^{-1} = \frac{n_d}{\sum_{j=1}^{n_d} w_{j|d}},$$

which calibrates the sum of the design weights within a domain to the known sample size. Alternatively, Pfeffermann et al. (1998) propose a scaling factor

$$\lambda_d = \frac{\sum_{j=1}^{n_d} w_{j|d}}{\sum_{j=1}^{n_d} w_{j|d}^2},$$

that can be interpreted as the inverse of a design effect.

Asparouhov (2006) gives a nice discussion on different options of how to scale the weights, including the mechanisms discussed by Pfeffermann et al. (1998). He uses simulation studies based on a simple random effects model to asses the behaviour of model estimates under different scaling options and for various degrees of informativeness. This includes a comparison of the procedures under invariant and non-invariant selection. Asparouhov (2006) defines a sampling process to be invariant if conditional on the model covariates the level one design weight $w_{j|d}$ is independent from the level two random effect v_d. He concludes that unweighted estimates may not exhibit large biases for the level two variance σ_v^2 under invariant selection but this does not hold for other selection procedures or level one variances (cf. Asparouhov, 2006). Interestingly, using unscaled weighted analysis performed even worse than unweighted estimation due to small sample biases. This issue is highly relevant for small area estimation and therefore unscaled weighted estimators should be avoided whenever possible.

The approach of Pfeffermann et al. (1998) focussed on linear mixed models. Extensions to generalised linear mixed models have been proposed by Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006), who maximise a design-weighted pseudo log-likelihood. A unifying approach maximising the weighted log-composite likelihood has been introduced by Rao, Verret, and Hidiroglou (2013).

While these procedures were mainly developed to support analytic inference, many strategies aiming to incorporate design weights in model-based small area predictors have been proposed as well. A very simple proposal in this regard is to transform the unit level model (3.2.1) to a survey-weighted area level model. Thereby, design weights are incorporated into the estimation of the fixed regression coefficients but not used when estimating the variance components. This approach has been proposed initially by Prasad and Rao (1999), who consider a random effects model without covariates, so that $\mathbf{x}_{dj} = 1$ for all units of the population. Thus, their unit level random effects model reads as follows:

$$y_{dj} = \mu + v_d + \varepsilon_{dj},\ j = 1, \ldots, N_d,\ d = 1, \ldots, D. \tag{4.2.1}$$

Since (4.2.1) is a special case of the nested error regression model (3.2.1), the assumptions on the random effects and the individual error terms remain unchanged, i.e. $v_d \overset{i.i.d.}{\sim} N(0, \sigma_v^2)$, $\varepsilon_{dj} \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2)$ and $\mathrm{Cov}(\varepsilon_{dj}, v_d) = 0\ \forall d, j$. In order to incorporate design weights, model (4.2.1) is transformed to a survey-weighted model on area level. To this end, a Hajèk-type estimator for the domain mean is constructed as

$$\overline{y}_{dw} = \frac{\sum_{j=1}^{n_d} w_{dj} y_{dj}}{\sum_{j=1}^{n_d} w_{dj}}.$$

Applying the same strategy on the right hand side of model (4.2.1) leads to the design-weighted area level model

$$\overline{y}_{dw} = \mu + v_d + \overline{\varepsilon}_{dw}, \tag{4.2.2}$$

with $\overline{\varepsilon}_{dw} = \dfrac{\sum_{j=1}^{n_d} w_{dj}\varepsilon_{dj}}{\sum_{j=1}^{n_d} w_{dj}}$. Due to the i.i.d. and normality assumptions on ε_{dw}, the error term in (4.2.2) is distributed as $\overline{\varepsilon}_{dw} \sim N(0, \sigma_\varepsilon^2 \sum_{j=1}^{n_d} \tilde{w}_{dj}^2)$, using $\tilde{w}_{dj} = \dfrac{w_{dj}}{\sum_{j=1}^{n_d} w_{dj}}$. Note that the random effect is unaltered compared to (4.2.1), since $\overline{v}_{dw} = \sum_{j=1}^{n_d} w_{dj} v_d / \sum_{j=1}^{n_d} w_{dj} = v_d \sum_{j=1}^{n_d} w_{dj} / \sum_{j=1}^{n_d} w_{dj} = v_d$, as v_d is the same for all units within domain d. The pseudo-BLUP minimising the prediction variance among the class of linear unbiased estimators under model (4.2.2) has been derived by Prasad and Rao (1999) but depends on the unknown model parameters $\boldsymbol{\xi} = (\mu, \sigma_v^2, \sigma_\varepsilon^2)^T$. Replacing $\boldsymbol{\xi}$ by model-consistent estimates $\widehat{\boldsymbol{\xi}}$ yields the pseudo-EBLUP, which is given by

$$\begin{aligned} \widehat{\mu}_d^{PR} &= \widehat{\mu}_w + \widehat{v}_{dw} \\ &= \widehat{\gamma}_{dw}\overline{y}_{dw} + (1 - \widehat{\gamma}_{dw})\widehat{\mu}_w, \end{aligned} \tag{4.2.3}$$

with $\widehat{\mu}_w = \sum_{d=1}^{D} \widehat{\gamma}_{dw}\overline{y}_{dw} / \sum_{d=1}^{D} \widehat{\gamma}_{dw}$, $\widehat{v}_{dw} = \widehat{\gamma}_{dw}(\overline{y}_{dw} - \widehat{\mu}_w)$ and $\widehat{\gamma}_{dw} = \widehat{\sigma}_v^2 / (\widehat{\sigma}_v^2 + \widehat{\sigma}_\varepsilon^2 \sum_{j=1}^{n_d} \tilde{w}_{dj}^2)$. Thus, the pseudo-EBLUP can be considered as a weighted average between the direct ratio estimator $\overline{y}_{dw}$ and the synthetic estimator $\widehat{\mu}_w$. Note that as with the BHF- and FH-model introduced in Section 3.2, the normality assumption is not needed to derive the pseudo-(E)BLUP (Prasad & Rao, 1999).

A problem which may arise when using (4.2.3) in practice, is the case of $\widehat{\sigma}_v^2 = 0$. This situation can occur if the variance components are estimated under the method of moments, which may lead to negative values for $\widehat{\sigma}_v^2$. As negative variance components are not within the admissible parameter space, the estimate $\widehat{\sigma}_v^2$ will be truncated to zero in this case. Assuming $\widehat{\sigma}_v^2 = 0$, it is easily seen that also $\widehat{\gamma}_{dw} = 0$ in all domains, implying that the pseudo-EBLUP reduces to the synthetic component $\widehat{\mu}_w$. Unfortunately, $\widehat{\mu}_w$ is not defined if $\widehat{\sigma}_v^2 = 0$, since $\sum_{d=1}^{D} \widehat{\gamma}_{dw} = 0$, which is the denominator of $\widehat{\mu}_w$. Following practical considerations, we could set $\widehat{\mu}_w = D^{-1} \sum_{d=1}^{D} \overline{y}_{dw}$ if $\widehat{\sigma}_v^2 = 0$.

An MSE estimator of (4.2.3) has been derived by Prasad and Rao (1999). It is given by

$$\begin{aligned} \widehat{\mathrm{MSE}}(\widehat{\mu}_d^{PR}) &= g_{1d}(\widehat{\boldsymbol{\delta}}) + g_{2d}(\widehat{\boldsymbol{\delta}}) + 2g_{3d}(\widehat{\boldsymbol{\delta}}) \quad \text{with} \\ g_{1d}(\widehat{\boldsymbol{\delta}}) &= (1 - \widehat{\gamma}_{dw})\widehat{\sigma}_v^2 \\ g_{2d}(\widehat{\boldsymbol{\delta}}) &= (1 - \widehat{\gamma}_{dw})^2 \widehat{\sigma}_v^2 / \sum_{d=1}^{D} \widehat{\gamma}_{dw} \\ g_{3d}(\widehat{\boldsymbol{\delta}}) &= \widehat{\gamma}_{dw}(1 - \widehat{\gamma}_{dw})^2 \widehat{\sigma}_v^{-2} \left[V(\widehat{\sigma}_v^2) - 2(\widehat{\sigma}_v^2/\widehat{\sigma}_\varepsilon^2)\mathrm{Cov}(\widehat{\sigma}_v^2, \widehat{\sigma}_\varepsilon^2) + (\widehat{\sigma}_v^2/\widehat{\sigma}_\varepsilon^2)^2 V(\widehat{\sigma}_\varepsilon^2) \right]. \end{aligned} \tag{4.2.4}$$

In (4.2.4), $V(\widehat{\sigma}_v^2)$, $V(\widehat{\sigma}_\varepsilon^2)$ and $\mathrm{Cov}(\widehat{\sigma}_v^2, \widehat{\sigma}_\varepsilon^2)$ denote the asymptotic variances of the domain-specific random effect, the individual random effect and the asymptotic covariance between both. These expressions depend on the method employed to estimate $\boldsymbol{\delta}$. Formulae under the methods of moments are given in Prasad and Rao (1999).

In their article, Prasad and Rao (1999) also provide an extension of the pseudo-EBLUP methodology to the case of nested error regression model (3.2.1) with covariates. The authors transform this model to a survey-weighted area level model given by

$$\overline{y}_{dw} = \overline{\mathbf{x}}_{dw}^T \boldsymbol{\beta} + v_d + \overline{\varepsilon}_{dw}, \tag{4.2.5}$$

where $\overline{\mathbf{x}}_{dw} = \sum_{j=1}^{n_d} w_{dj}\mathbf{x}_{dj} / \sum_{j=1}^{n_d} w_{dj} = \sum_{j=1}^{n_d} \widetilde{w}_{dj}\mathbf{x}_{dj}$. The pseudo-EBLUP under model (4.2.5) follows as (cf. You & Rao, 2002a)

$$\begin{aligned} \widehat{\mu}_d^{PR2} &= \overline{\mathbf{X}}_d^T \widehat{\boldsymbol{\beta}}_w + \widehat{\gamma}_{dw}(\overline{y}_{dw} - \overline{\mathbf{x}}_{dw}^T \widehat{\boldsymbol{\beta}}_w) \quad \text{where} \\ \widehat{\boldsymbol{\beta}}_w &= (\widehat{\gamma}_{dw}\overline{\mathbf{x}}_{dw}\overline{\mathbf{x}}_{dw}^T)^{-1}(\widehat{\gamma}_{dw}\overline{\mathbf{x}}_{dw}\overline{y}_{dw}). \end{aligned} \tag{4.2.6}$$

To apply estimator (4.2.6), first the variance components $\boldsymbol{\delta}$ are estimated from the available unit level survey data. Prasad and Rao (1999) considered the method of moments to estimate $\boldsymbol{\delta}$ but other methods such as ML or REML can be used as well. Then, in a second step, we can estimate $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}_w$ from the aggregated area level data for given values of $\widehat{\boldsymbol{\delta}}$. It might be noted that the pseudo-EBLUP (4.2.6) is similar to the EBLUP under the area level model given by (3.2.21). Both estimators are design-consistent and use area level information to obtain the vector of regression coefficients. While under the Fay-Herriot model, an estimate of $\boldsymbol{\beta}$ is obtained from regressing the direct estimates on known auxiliary domain information, the pseudo-EBLUP approach uses a regression from the direct estimates on estimated auxiliary domain information. Furthermore, the variance components are obtained using unit level data in the pseudo-EBLUP, whereas in the Fay-Herriot model known variances of the direct estimates are assumed and the variance of the linking model is estimated using area level information only.

One issue with (4.2.6) is that the vector of regression parameters is estimated using aggregate data only, while unit level information are available. This may be unsatisfactory, since efficiency losses may occur. Alternatively, as noted by You and Rao (2002a), the BLUP of v_d can be expressed in terms of the aggregated area level model as

$$\widetilde{v}_{dw} = \gamma_{dw}\left(\overline{y}_{dw} - \overline{\mathbf{x}}_{dw}^T \boldsymbol{\beta}\right). \tag{4.2.7}$$

Now, $\widetilde{v}_{dw} = \widetilde{v}_{dw}(\boldsymbol{\beta}, \boldsymbol{\delta})$ is a function of the model parameters, as it depends on the true vector of regression parameters and the true variance components. Plugging this expression into the survey-weighted estimating equations gives:

$$\sum_{d=1}^{D}\sum_{j=1}^{n_d} w_{dj}\mathbf{x}_{dj}\left(y_{dj} - \mathbf{x}_{dj}^T\boldsymbol{\beta} - \gamma_{dw}\left(\overline{y}_{dw} - \overline{\mathbf{x}}_{dw}^T\boldsymbol{\beta}\right)\right) = 0. \tag{4.2.8}$$

Rearranging (4.2.8) leads to

$$\sum_{d=1}^{D}\sum_{j=1}^{n_d} w_{dj}\mathbf{x}_{dj}\left(\mathbf{x}_{dj}-\gamma_{dw}\overline{\mathbf{x}}_{dw}\right)^T\boldsymbol{\beta} = \sum_{d=1}^{D}\sum_{j=1}^{n_d} w_{dj}\mathbf{x}_{dj}\left(y_{dj}-\gamma_{dw}\overline{y}_{dw}\right) \tag{4.2.9}$$
$$= \sum_{d=1}^{D}\sum_{j=1}^{n_d} w_{dj}\left(\mathbf{x}_{dj}-\gamma_{dw}\overline{\mathbf{x}}_{dw}\right)\overline{y}_{dw}.$$

Solving (4.2.9) for $\boldsymbol{\beta}$ yields an estimator of the regression coefficients (cf. You & Rao, 2002a):

$$\widetilde{\boldsymbol{\beta}}_{YR} = \sum_{d=1}^{D}\sum_{j=1}^{n_d}\left(w_{dj}\mathbf{x}_{dj}\left(\mathbf{x}_{dj}-\gamma_{dw}\overline{\mathbf{x}}_{dw}\right)^T\right)^{-1}\left(\sum_{d=1}^{D}\sum_{j=1}^{n_d} w_{dj}\left(\mathbf{x}_{dj}-\gamma_{dw}\overline{\mathbf{x}}_{dw}\right)\overline{y}_{dw}\right). \tag{4.2.10}$$

It has been noted by Huang and Hidiroglou (2003) that there is a strong similarity between (4.2.10) and the procedure due to Pfeffermann et al. (1998). They observed that when the latter is adjusted for SRS, expression (4.2.10) results.

In reality, the variance components are not known but have to be estimated from the data. Since $\widetilde{\boldsymbol{\beta}}_{YR}$ depends on $\boldsymbol{\delta}$ as well, we have to replace it by $\widehat{\boldsymbol{\beta}}_{YR} = \widetilde{\boldsymbol{\beta}}_{YR}(\widehat{\boldsymbol{\delta}})$. The pseudo-EBLUP using unit level data for estimating $\boldsymbol{\beta}$ has the same structure as (4.2.6) and follows as (You & Rao, 2002a):

$$\widehat{\mu}_d^{YR} = \overline{\mathbf{X}}_d^T\widehat{\boldsymbol{\beta}}_{YR} + \widehat{\gamma}_{dw}(\overline{y}_{dw} - \overline{\mathbf{x}}_{dw}^T\widehat{\boldsymbol{\beta}}_{YR}). \tag{4.2.11}$$

In addition to achieving design consistency, the You-Rao estimator (4.2.11) is coherent with design-based estimates, as it is benchmarked against a GREG estimator of the national total. This property of the You-Rao estimator was derived by You and Rao (2002a). Since a direct GREG estimator is commonly used for estimating national quantities, it is beneficial if the model-based small area estimates are coherent with the national figure. If the estimates are required on various level of aggregation, however, the You-Rao estimates are not benchmarked against intermediate levels. To estimate the MSE of (4.2.11), You and Rao (2002a) apply the methodology of Prasad and Rao (1990), which yields

$$\begin{aligned}
\widehat{\mathrm{MSE}}(\widehat{\mu}_d^{YR}) &= g_{1d}(\widehat{\boldsymbol{\delta}}) + g_{2d}(\widehat{\boldsymbol{\delta}}) + 2g_{3d}(\widehat{\boldsymbol{\delta}}) \quad \text{with} \\
g_{1d}(\widehat{\boldsymbol{\delta}}) &= (1-\widehat{\gamma}_{dw})\widehat{\sigma}_v^2 \\
g_{2d}(\widehat{\boldsymbol{\delta}}) &= \left(\overline{\mathbf{X}}_d - \widehat{\gamma}_{dw}\overline{\mathbf{x}}_{dw}\right)^T\Phi_w\left(\overline{\mathbf{X}}_d - \widehat{\gamma}_{dw}\overline{\mathbf{x}}_{dw}\right) \\
g_{3d}(\widehat{\boldsymbol{\delta}}) &= \widehat{\gamma}_{dw}(1-\widehat{\gamma}_{dw})^2\widehat{\sigma}_v^{-2}\widehat{\sigma}_\varepsilon^{-4}\left[\widehat{\sigma}_\varepsilon^4 V(\widehat{\sigma}_v^2) - 2\widehat{\sigma}_v^2\widehat{\sigma}_\varepsilon^2\mathrm{Cov}(\widehat{\sigma}_v^2,\widehat{\sigma}_\varepsilon^2) + \widehat{\sigma}_v^4 V(\widehat{\sigma}_e^2)\right].
\end{aligned} \tag{4.2.12}$$

In (4.2.12) Φ_w denotes the variance-covariance matrix of $\widetilde{\boldsymbol{\beta}}_{YR}$, which is given in equation 16 of You and Rao (2002a). As noted by Torabi and Rao (2010), the MSE estimator (4.2.12) is nearly unbiased in the case of a sampling design which is self-weighting within areas. For other sampling designs, however, the MSE estimator (4.2.12) is no longer nearly unbiased, since cross-product terms are ignored in expression (4.2.12). A nearly unbiased

MSE estimator which accounts for these cross-product terms has been derived by Torabi and Rao (2010).

More generally, the pseudo-EBLUP approach due to Prasad and Rao (1999) can be extended to model-assisted empirical best prediction approaches. This class of estimators was introduced by Jiang and Lahiri (2006a) and is not restricted to linear mixed models. Their procedure yields an estimator which is both design- and model-consistent. Moreover, the model-assisted EBP does not rely on the implicit assumption that the survey-weighted area level model holds also for the non-sampled part of the population (Jiang & Lahiri, 2006a, p. 303).

It should be noted that the primary purpose of the approaches due to Prasad and Rao (1999) and You and Rao (2002a) is to yield estimators which are both design-consistent and robust against model misspecification. However, studies have indicated their potential to compensate as well to some extent for the effects of informative sampling (cf. Verret et al., 2015 and references therein). As noted by Pfeffermann (2002, p. 138f.), these estimators protect against informative sampling within areas but do not account for the potentially informative selection of areas. Moreover, he argues that in the case of a non-ignorable selection within areas, the variances should be estimated under the randomisation distribution with respect to the design (Pfeffermann, 2006, p. 70f.).

4.2.2 Using design information as auxiliary information

An alternative approach is to include among the covariates all variables which determine the selection of the sample such that conditional on the covariates the selection mechanism can be ignored. We will illustrate this strategy for the case of the random intercept model (3.2.1). Denoting the additional covariates by $\mathbf{z}_{dj}$, the model which is fitted to the sample data is given by

$$y_{dj} = \mathbf{x}_{dj}^T\boldsymbol{\beta} + \mathbf{z}_{dj}^T\boldsymbol{\kappa} + v_d + \varepsilon_{dj}, \quad j = 1, \ldots, n_d, \quad d = 1, \ldots, D, \tag{4.2.13}$$

where $\boldsymbol{\kappa}$ is the vector of regression effects associated with the design variables $\mathbf{z}_{dj}$ and the assumptions on v_d and ε_{dj} remain the same as in (3.2.1). As $\mathbf{z}_{dj}^T$ comprises all variables that determine the sample selection,

$$\begin{aligned} \Pr(I_d = 1|v_d) &= \Pr(I_d = 1) \\ \Pr(I_{dj} = 1|y_{dj}, \mathbf{x}_{dj}, v_d, I_d = 1, \mathbf{z}_{dj}) &= \Pr(I_{dj} = 1|\mathbf{x}_{dj}, v_d, I_d = 1, \mathbf{z}_{dj}), \end{aligned} \tag{4.2.14}$$

are satisfied and the extended model holds for the population as well (cf. Rao, 2003, Selection 5.3). Defining $\mathbf{x}_{dj}^{*T} := (\mathbf{x}_{dj}^T, \mathbf{z}_{dj}^T)$ and $\boldsymbol{\beta}^{*T} := (\boldsymbol{\beta}^T, \boldsymbol{\kappa}^T)$, we can rewrite the extended model as

$$y_{dj} = \mathbf{x}_{dj}^{*T}\boldsymbol{\beta}^* + v_d + \varepsilon_{dj}, \quad j = 1, \ldots, n_d, \quad d = 1, \ldots, D. \tag{4.2.15}$$

Since (4.2.15) has the same structure as (3.2.1), the same model-based predictors as discussed in Section 3.2 can be used. The only change is that the model is now richer. It should be noted, however, that using an extended model requires knowledge of the population means of $\overline{\mathbf{Z}}_d$, to produce EBLUP estimates as described in Section 3.2. This approach is not feasible if the sample selection depends on the outcome variable, e.g. as in

endogenous stratification (cf. Pfeffermann & Sverchkov, 2009, p. 463). But if the sample selection does not depend on the outcome it is a very efficient strategy. Moreover, this approach is very convenient in finite population sampling, as the aim is to get unbiased predictions with a small prediction error. In the case of a business survey where the sampling design amounts to a StrRS, this implies the inclusion of the stratum-defining variables in the model, e.g. the industry classification, macro-region and employee size class. Since these variables can be safely assumed to be known by the national statistical institute conducting the survey, it will be pretty easy to include them in the modelling. This claim may be overly optimistic when the model analyst does not have the necessary information about design variables, which may occur due to confidentiality reasons. Besides, if the aim of the analysis is to learn about the model parameters and to attribute a causal interpretation on some of these coefficients, including the design information may not be the best choice. This issue is very important in analytic inference. However, finite population sampling and thus small area estimation belong the field of descriptive inference, where the model is used as a tool to derive predictions for the non-sampled units. Another issue, which might be more relevant for small area estimation is that this approach might lead to very complex models, which can be difficult to fit (cf. Pfeffermann & Sverchkov, 2009, p. 462). This may pose challenges especially for complex non-linear unit level models. One issue that could arise in this respect is an ill-conditioned design matrix $\mathbf{X}^*$ composed of the sample units of $\mathbf{x}_{dj}^{*T}$, which could lead to problems when estimating $\boldsymbol{\beta}^*$.

If the model analyst has access to the design weights w_k for all units in the population, the above problem can be simplified tremendously. This is due to the following result established by Skinner (1994) in the context of a linear fixed effects model:

$$f_S(y_k|\mathbf{x}_k, w_k) = f_U(y_k|\mathbf{x}_k, w_k). \tag{4.2.16}$$

An important prerequisite to derive (4.2.16) is the random indexing assumption, which requires that the $y_k, \mathbf{z}_k$ and $\mathbf{x}_k$ are jointly independent and identically distributed. This and the fact that the $w_k = \pi_k^{-1}$ are a sufficient statistic for $I_k = 1$ yield (4.2.16) (Chambers et al., 2003; Skinner, 1994). Equation (4.2.16) states that the marginal model of y_k given the covariates $\mathbf{x}_k$ and w_k is the same for the sample and the population. Hence, the inclusion probabilities π_k provide an adequate summary of the design variables $\mathbf{z}_k$ used in the sample selection process and the sampling design is not informative (cf. Pfeffermann & Sverchkov, 2009, p. 463). Note that this refers to a marginal sample model, which is in contrast to the joint models considered in much of the literature (see for example the definitions of an informative sampling design in Pfeffermann & Sverchkov, 2009, Section 2). To apply Skinner's result to small area estimation under the mixed model (3.2.1), we need to account for the two-level structure and need area-specific versions of (4.2.16), as only within areas conditional on the random effect, the i.i.d. assumption can be maintained. Following the literature on informative sampling and small area estimation, we assume the two-stage design outlined in Algorithm 2. Hence, the inverse of the conditional second-stage inclusion probabilities will be denoted as $\pi_{j|d}^{-1} = w_{j|d}$. It may be noted that instead of $w_{j|d}$ any function of the selection probabilities $g(p_{j|d})$ could be included as an additional covariate, leading to so-called augmented estimation methods pioneered by Verret et al. (2015). Formally, this procedure can be viewed as a special case of the extended model (4.2.13), where $\mathbf{z}_{dj}$ reduces to a scalar given by $z_{dj} = g(p_{j|d})$. In the most recent version of their paper, Verret et al. (2015) consider various specifications of $g(p_{j|d})$. Ideally, the choice of $g(\cdot)$ should depend on the structure of the informative design. Verret et al. (2015) propose to look at a plot of the composite residuals against the

various choices to $g(p_{j|d})$ to decide on the functional form of $g(\cdot)$. If a linear relationship between the composite residuals and a particular function $g(\cdot)$ can be detected in a sample, than this augmentation strategy should work well. The plot of residuals against omitted variables is a well-known technique in regression analysis to check whether important covariates are missing in the model specification. Hence, augmented estimation can be seen as a technique to avoid model misspecification. Another potential strategy to decide on whether augmentation should be employed and which kind should be applied if any, is to resort to the model selection tools introduced in Section 3.5. In this regard, the conditional AIC (3.5.2) due to Vaida and Blanchard (2005), provides a natural tool to decide on the specification of the augmented model. In the simulation study of Verret et al. (2015), using the baseline design weight $w_{j|d}$ as the additional covariate was clearly dominated by other augmentation strategies, such as choosing $g(\cdot)$ to be the identity, inverse or log. As noted by Verret et al. (2015) this finding could be due to their setup, where $w_{j|d}$ contains a term which does not relate to the informativeness of the sampling design. In practice, the model analysts' choice of $g(\cdot)$ can be restricted, due to data availability limitations. Obviously, augmentation is a much simpler strategy, especially from a computational viewpoint, compared to including all design variables in the model. Moreover, Verret et al. (2015) consider a pseudo-EBLUP under the augmented model, which enjoys the benchmarking property without further ado. Their simulation results indicate that there are circumstances under which a pseudo-EBLUP under the augmented model will yield better results than the EBLUP under the augmented model.

4.2.3 Modelling the sample selection process

The two preceding sections discussed strategies to overcome informative sampling designs either by using a design-consistent estimator or by augmenting the vector of covariates. Alternatively, the sample selection process may be modelled directly, thereby avoiding the need to know some design variables for all units in the population or to cope with efficiency losses. The idea to model the sample selection process is not exclusively used in frequentist model-based finite population sampling, but has been used extensively in econometrics (cf. Heckman, 1979) and in Bayesian statistics (cf. Little, 2003) as well. Pfeffermann and Sverchkov (2007) assume that the random intercept model (3.2.1) holds for the sampled units, which were selected using the two-stage design given in Algorithm 2. Furthermore, they assume that weights $w_{j|d}$ in sampled areas satisfy

$$\mathrm{E}_S(w_{j|d}|y_{dj}, \mathbf{x}_{dj}, v_d, I_d = 1) = k_d \exp(\mathbf{x}_{dj}^T \mathbf{a} + b y_{dj}), \tag{4.2.17}$$

where $k_d \propto N_d/n_d$ for large areas (Pfeffermann & Sverchkov, 2007, p. 1430). In the case that sampling fractions within selected areas are small, the authors derive the MSE-optimal predictor correcting for the informative design given by

$$\widehat{\mu}_d^{PS} = \overline{\mathbf{X}}_d^T \boldsymbol{\beta} + \widetilde{v}_d + \left(1 - \frac{n_d}{N_d}\right) b\sigma_\varepsilon^2 \tag{4.2.18}$$

for sampled areas and

$$\widehat{\mu}_d^{PS} = \overline{\mathbf{X}}_d^T \boldsymbol{\beta} + b\sigma_\varepsilon^2 + \frac{\sum_{d\in S}(w_d - 1)\widetilde{v}_d}{\sum_{d\in S}(w_d - 1)} \tag{4.2.19}$$

for non-sampled areas. In (4.2.19), the last term corrects for the informativeness of selection of areas. It can be seen from both equations that informativeness of the sampling within areas is accounted for by the term $b\sigma_\varepsilon^2$. Of course, neither (4.2.18) nor (4.2.19) are applicable in practice, as they depend on unknown model parameters. Hence the parameters $\boldsymbol{\beta}$, σ_v^2, σ_ε^2 and b have to be replaced by suitable estimates. To estimate $\boldsymbol{\beta}$, σ_v^2 and σ_ε^2, the same methods as described in previous sections are applicable. As b depends on the non-linear weighting model (4.2.17), we need to estimate this model. As pointed out by Verret et al. (2015), starting values for the non-linear least squares algorithm can be obtained from regressing $\log(w_{j|d})$ on $\mathbf{x}_{dj}$ and y_{dj}. To estimate the MSE of their predictor, Pfeffermann and Sverchkov (2007) propose to use a parametric bootstrap.

4.3 Incorporating the design under a log-transformed unit level model

The discussion in the previous section highlighted a variety of approaches to include the sampling design into model-based small area procedures under a linear mixed model. Now, we are going to discuss how some of these ideas could be translated to the case of the lognormal mixed model (3.4.5). This issue seems important as the EBP (3.4.10) under this model does not make use of design weights. Hence, it is not design-consistent and may suffer from severe biases in the case of an informative sampling scheme. Note that the model-based direct estimator proposed by Chandra and Chambers (2011) does not solve the problem either, although the weighted sample sum of the fitted values reproduces the sum of the fitted values in the population. This is due to the fact that under informative sampling, the fitted values for the non-sampled units are biased predictions of the variable of interest. Hence, the weights calibrate against a biased prediction of the total.

4.3.1 Augmented estimation

One option to account for the potential bias due a non-ignorable sampling design is by including all the design variables $\mathbf{z}_{dj}$ into the model, just as in the case of the nested error regression model. This yields a model

$$\log(y_{dj}) = \boldsymbol{x}_{dj}^T\boldsymbol{\beta} + \mathbf{z}_{dj}^T\boldsymbol{\kappa} + v_d + e_{dj}, \quad d = 1, \ldots, D,\ j = 1, \ldots, N_d, \tag{4.3.1}$$

with the same assumptions on v_d and e_{dj} as in (3.4.5). Since all design information is included in $\mathbf{z}_{dj}$, the sample selection does not depend on y_{dj} conditional on $\mathbf{z}_{dj}$ and thus the model holds for both the sample and the population. It should be noted, however, that owing to the non-linear back-transformation present in the BP, using (4.3.1) requires the knowledge of $\mathbf{z}_{dj}$ for all non-sampled elements as well. In the previous section, the issue has been raised that knowledge about all the design information for all units of the population may not be a realistic option in practice. Instead, we apply the strategy of augmentation proposed by Verret et al. (2015) in the context of the linear mixed model, where a function of the selection probabilities is the only design information used in the modelling, i.e. $\mathbf{z}_{dj} = g(p_{j|d})$ and hence also $\boldsymbol{\kappa} \equiv \kappa$. Thus, model (4.3.1) simplifies to

$$\log(y_{dj}) = \boldsymbol{x}_{dj}^T\boldsymbol{\beta} + g(p_{j|d})\kappa + v_d + e_{dj}, \quad d = 1, \ldots, D,\ j = 1, \ldots, N_d. \tag{4.3.2}$$

This leads to the BP under the augmented model given by

$$\widetilde{\mu}_d^{Aug} = \frac{1}{N_d}\left[\sum_{j\in S_d} y_{dj} + \sum_{j\notin S_d} \widetilde{y}_{dj}^{Aug}\right], \quad d = 1, \ldots, D \quad \text{where} \tag{4.3.3}$$

$$\widetilde{y}_{dj}^{Aug} = \exp(\mathbf{x}_{dj}^T\boldsymbol{\beta} + \kappa g(p_{j|d}) + \widetilde{v}_d + 0.5\sigma_\varepsilon^2(\gamma_d/n_d + 1)). \tag{4.3.4}$$

which depends on the unknown parameters $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \kappa, \sigma_v^2, \sigma_\varepsilon^2)^T$ and hence cannot be computed in practice. Replacing the unknown model parameters by model-consistent estimates $\widehat{\boldsymbol{\xi}}$ in equation (4.3.3) yields the EBP as

$$\widehat{\mu}_d^{Aug} = \frac{1}{N_d}\left[\sum_{j\in S_d} y_{dj} + \sum_{j\notin S_d} \widehat{y}_{dj}^{Aug}\right], \quad d = 1, \ldots, D \quad \text{where} \tag{4.3.5}$$

$$\widehat{y}_{dj}^{Aug} = \exp(\mathbf{x}_{dj}^T\widehat{\boldsymbol{\beta}} + \widehat{\kappa} g(p_{j|d}) + \widehat{v}_d + 0.5\widehat{\sigma}_\varepsilon^2(\widehat{\gamma}_d/n_d + 1)). \tag{4.3.6}$$

As a comparison between this EBP using the augmented model and the EBP under ignorable sampling (3.4.10) shows, they differ due to the presence of the $\widehat{\kappa}g(p_{j|d})$-term, which accounts for the informativeness. To decide between (4.3.5) and (3.4.10), the (conditional) AIC of both underlying models can be used, as it is a measure of the predictive accuracy. Note that the EBP under the augmented model requires knowledge of $g(p_{j|d})$ for all units in the population, which may be difficult to obtain for the model analyst with πps sampling. Nonetheless, in the case of StrRS where all strata are sampled, the requirement is less likely to be a huge impediment. The MSE of $\widehat{\mu}_d^{Aug}$ may be estimated using the same strategies as for the EBP that have been presented in Section 3.4, because the EBP under the augmented model has the same structure as the EBP under the model without augmentation.

4.3.2 Using weighted model parameter estimates

As an alternative approach, we may want to obtain design-weighted estimates of the model parameters $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \sigma_v^2, \sigma_\varepsilon^2)$. This can be done in the spirit of You and Rao (2002a), by employing survey weighted estimating equations to estimate $\boldsymbol{\beta}$. You and Rao (2002a) observed that conditional on the variance components, the resulting estimator is model-unbiased for $\boldsymbol{\beta}$. It should be noted that the resulting estimates no longer satisfy the benchmarking property (3.6.7) without further adjustment. Besides, also replicating the survey at hand is an option. Such an approach was described by Huang and Hidiroglou (2003) for linear mixed models. The idea is that the design weight indicates the number of population units represented by each sampled unit. Assuming integer weights, we could then replicate each entry in the sample times the corresponding weight.

Alternatively, we might want to consider the model-assisted EBP due to Jiang and Lahiri (2006a), which does not require to model the non-sampled part of the population. However, their approach requires to specify the conditional density of $\overline{y}_{dw}$ given v_d. Under the

lognormal mixed model $\overline{y}_{dw}|v_d$ amounts to a weighted sum of conditionally independent lognormal random variables, whose density does not have a closed-form expression (cf. Kleiber & Kotz, 2003, p. 111). Hence, to apply the method of Jiang and Lahiri (2006a), we would need to resort to approximations of this conditional density.

4.4 Simulation studies

We compare the different strategies to account for the sampling design in lognormal mixed models for small area estimation by means of simulation studies. We use the same population structure as Verret et al. (2015), i.e. we consider $D = 99$ domains, each consisting of $N_d = 100$ elements. We follow Verret et al. (2015), as we apply sampling with probabilities proportional to size within all domains. Thus, we do not have to distinguish between inclusion probabilities for domains and for units within domains, such that the latter will be denoted as π_{dj} for simplicity. This yields inclusion probabilities as

$$\pi_{dj} = n_d b_{dj} / \sum_{j=1}^{N_d} b_{dj} = n_d p_{dj}, \quad d = 1, \ldots, D,\ j = 1, \ldots, N_d, \tag{4.4.1}$$

where b_{dj} is the size variable used. Thus, the inclusion probabilities result as the domain-specific sample size multiplied by selection probabilities $p_{dj} = b_{dj} / \sum_{j=1}^{N_d} b_{dj}$. It should be noted that the π_{dj} are the final inclusion probabilities and that they sum to n_d in every domain. The samples were drawn by means of Midzuno's method which is described in detail in (Tillé, 2006, Section 6.3.5). We chose Midzuno's method as a fast C++ implementation is available in the R package `simFrame` (Alfons, Templ, & Filzmoser, 2010). Regarding the choice of the size measure, we follow Verret et al. (2015) and consider a size variable due to Pfeffermann and Sverchkov (2007) as well as a size variable introduced by Asparouhov (2006) under invariant selection. In the following, we refer to the former approach as the PS size measure, whereas the latter will be called the Asparouhov size measure. As Verret et al. (2015), we set the domain-specific sample sizes to 5 for the first 33 areas, 7 for the next 33 areas and 9 for the remaining 33 areas. The model which was used in generating the populations is identical to parameter configuration 1 in Berg and Chandra (2014). Their model is given by

$$\log(y_{dj}) = -1.62 + 0.9x_{dj} + v_d + \varepsilon_{dj}, \quad d = 1, \ldots, D,\ j = 1, \ldots, N_d, \tag{4.4.2}$$

with $v_d \overset{i.i.d.}{\sim} N(0, 0.55^2)$ and $\varepsilon_{dj} \overset{i.i.d.}{\sim} N(0, 0.55^2/0.51)$. Furthermore, x_{dj} was drawn once as $x_{dj} \overset{i.i.d.}{\sim} N(3.253, 1.58^2)$ and held fixed over the simulation study. Note that this differs from the setting in Berg and Chandra (2014), who draw new realisations of x_{dj} in each simulation run.

An overview of the different estimators analysed in this study is given in Table 4.2. We performed the MSE estimation by means of a parametric bootstrap approach using 499 bootstrap replications for the estimators under study.

Table 4.2: Estimators in study on informative sampling under a lognormal mixed model

Abbreviation	Description
EBP	EBP defined in (3.4.10)
Augmented	EBP under augmented model defined in (4.3.5)
SWEE	Like (3.4.10), but $\widehat{\boldsymbol{\beta}}$ obtained by solving survey-weighted estimating equations

4.4.1 Pfeffermann-Sverchkov size measure

As a first scenario, we assume that the size variable considered by Pfeffermann and Sverchkov (2007) is used. This leads to the following size variable

$$b_{dj} = \exp\left((-(v_d + \varepsilon_{dj})/\sigma_\varepsilon + a_{dj}/5)/3\right), \tag{4.4.3}$$

where $a_{dj} \overset{i.i.d.}{\sim} N(0,1)$. The effects of using this particular design are illustrated in Figure 4.1 for one randomly selected sample. The blue line depicts the estimated density function for $\log(y_{dj})$ for the non-sampled units, whereas the magenta line shows the respective density function for the sampled values in this replication. It can be seen that the sample distribution is shifted to the left. A similar plot for the auxiliary variable x_{dj}, depicted in Figure 4.2, does not indicate systematic deviations between the sampled and non-sampled units. Hence, the sample model of $\log(y_{dj})|x_{dj}$ differs from the one for the non-sampled units, thereby confirming the informativeness of the sample.

We rely on a variety of diagnostics to select the augmenting variable. The first measure we consider was also employed by Verret et al. (2015, Section 4.2) and depicts the residuals from a regression

$$\log(y_{dj}) = \beta_0 + \beta_1 x_{dj} + u_{dj}, \tag{4.4.4}$$

where $u_{dj} = v_d + \varepsilon_{dj}$, estimated by OLS against potential augmenting variables. Ideally, a linear trend can be detected in this scatterplot, indicating the usefulness of the missing covariate to compensate for an omitted variable bias (cf. Verret et al., 2015). Figure 4.3 displays these plots for one particular sample. It can be seen that the design weight w_{dj} is not the best choice, because its relationship with the $\widehat{u}_{dj}$ is clearly non-linear. Moreover, the residual plots versus the selection probabilities p_{dj} and the inverse selection probabilities p_{dj}^{-1} both indicate informativeness, but they display a non-linear trend. Thus, using $\log(p_{dj})$ is the preferred choice as this plot reveals a linear pattern. Besides, we inspect the predictive accuracy of the different choices by calculating the cAIC. We choose the conditional over the marginal AIC, because the sampling design rules out non-sampled domains. Hence, we are interested in the predictive accuracy conditional on the vector of random effects. Table 4.3 reports the cAIC for the various model choices in one sample. The comparison of the cAIC values confirms the findings of the residual plots in Figure 4.3. Including the design weights leads already to strong improvement of the predictive accuracy of the model compared to the non-augmented model. However, the other choices are preferable among which $\log(p_{dj})$ clearly yields the highest predictive accuracy.

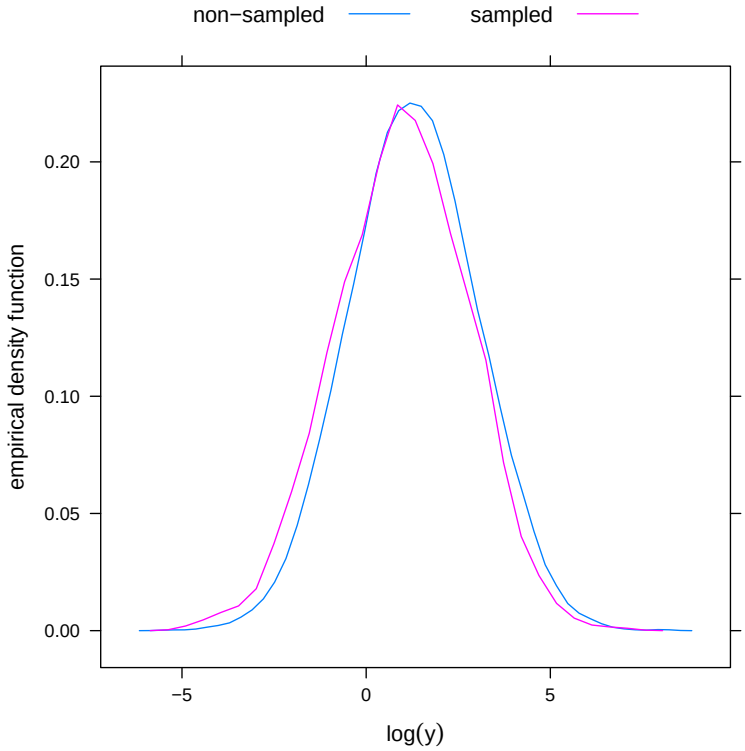

Figure 4.1: Empirical density functions of $\log(y_{dj})$ for sampled and non-sampled units under informative sampling with invariant selection

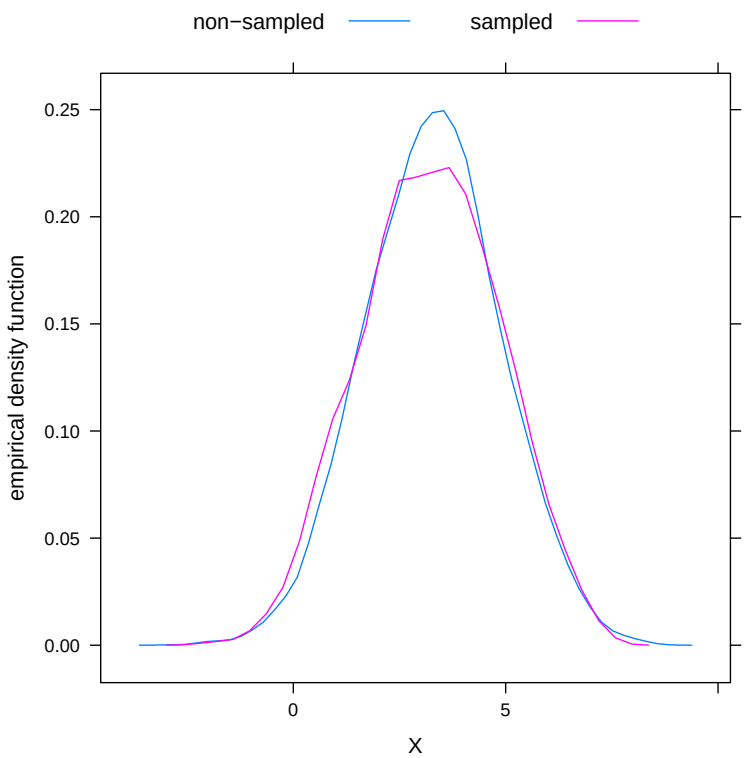

Figure 4.2: Empirical density functions of x_{dj} for sampled and non-sampled units under informative sampling with invariant selection

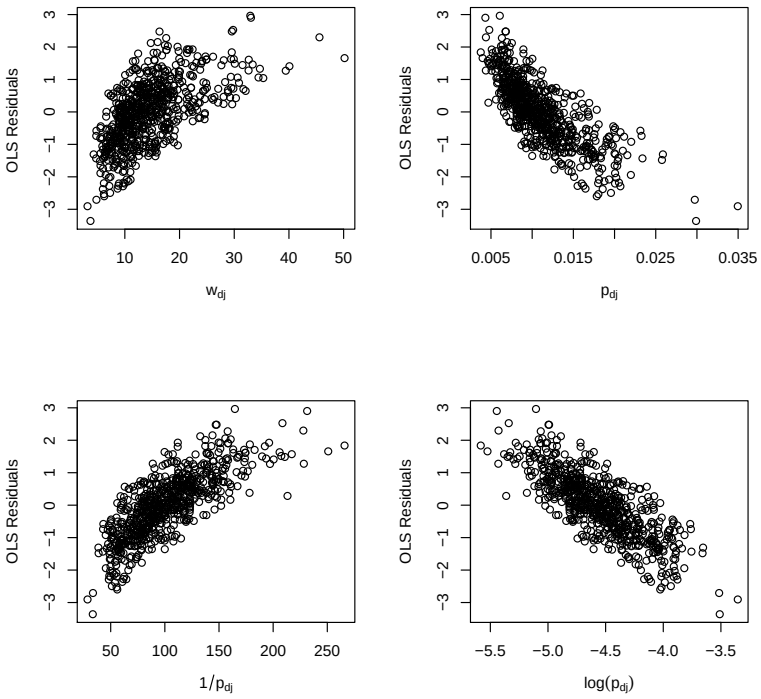

Figure 4.3: OLS residuals plotted against various choices of the augmenting variable under the PS size measure

Table 4.3: cAIC for augmented modelling under the PS size measure

	Augmented variable				
	none	w_{dj}	p_{dj}	p_{dj}^{-1}	$\log(p_{dj})$
$\log g(\mathbf{y}\|\widehat{\boldsymbol{\xi}}(\mathbf{y}), \widehat{\mathbf{v}}(\mathbf{y}))$	-811.77	-98.32	44.71	44.91	373.77
K	76.27	100.02	99.44	99.38	100.98
cAIC	1776.08	396.67	109.45	108.95	-545.58

In addition to the residual plots and the predictive accuracy, we propose to investigate the normality assumptions on v_d and ε_{dj} as well. These assumptions are crucially important under the lognormal mixed model as the expression for the EBP explicitly uses properties of the lognormal distribution (cf. Section 3.4 of this thesis). For a joint assessment of the normality assumptions, we consider QQ plots of the transformed residuals given by:

$$\widetilde{e}_{dj} = \left(\log(y_{dj}) - \widehat{\alpha}_d \bar{l}_d\right) - \left(\mathbf{x}_{dj} - \widehat{\alpha}_d \overline{\mathbf{x}}_d\right)^T \widehat{\boldsymbol{\beta}}, \tag{4.4.5}$$

where $\bar{l}_d = n_d^{-1} \sum_{j=1}^{n_d} \log(y_{dj})$ and $\widehat{\alpha}_d = 1 - (1 - \widehat{\gamma}_d)^{1/2}$. The QQ plots of the transformed residuals under various augmentation strategies for the particular sample are displayed in Figure 4.4. While we do not see any evidence of non-normality when $\log(p_{dj})$ is used, the plots with w_{dj}, p_{dj} and p_{dj}^{-1} display some non-linearities at the tails. After inspecting all these diagnostics, we decide to focus on $\log(p_{dj})$ as the augmenting variable, since the model including it satisfies the normality assumptions and yields the highest predictive

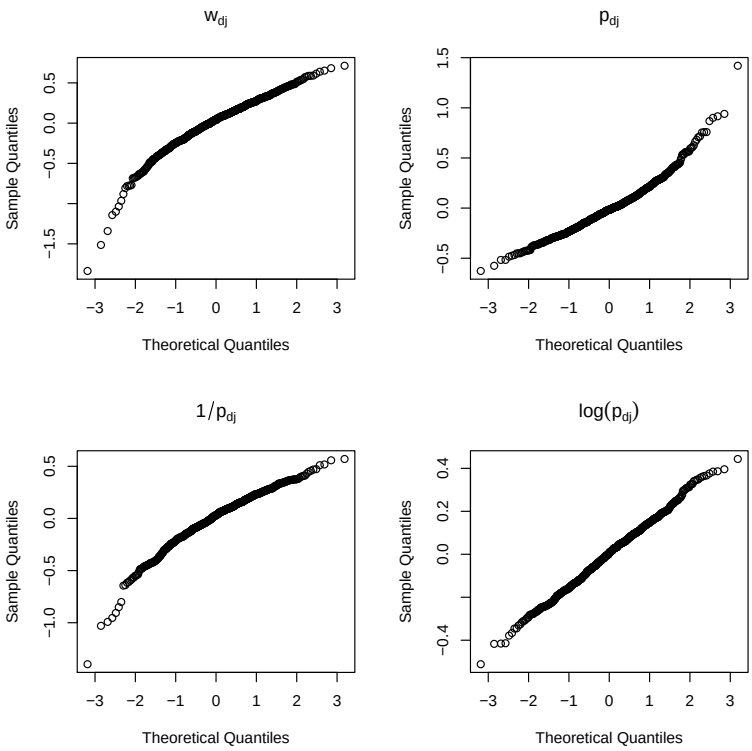

Figure 4.4: QQ plots of the transformed residuals for various choices of the augmenting variable under the PS size measure

accuracy among the choices considered.

The relative biases as well as the relative root mean squared errors (RRMSEs) under the PS size measure (4.4.3) are depicted in Figure 4.5. It is easily seen from the left panel of Figure 4.5 that the EBP without augmentation exhibits severe biases under the PS size measure. The SWEE predictor achieves slightly better results compared to the EBP but is still negatively biased. Unbiased point estimates can be achieved when using the EBP under the augmented model. Moreover, the results in terms of the RRMSE confirm the superiority of the point estimates obtained by augmentation, as this method yields the most precise estimates with RRMSEs below 10 per cent for most of the domains. Regarding the other two estimators, the SWEE performs better than the plain EBP. However, the smallest RRMSE achieved by the SWEE predictor is about 30 per cent and thus far worse than the worst RRMSE achieved by the EBP under the augmented model

To study the quality of the precision estimates as well, the confidence interval coverage rates are plotted against the average length of the intervals. This plot is shown in Figure 4.6, where the red line in each panel indicates the nominal coverage rate of 95 per cent. It may be noted that while both the EBP under the augmented model and the SWEE estimator achieve the nominal coverage rate, the EBP without augmentation suffers from undercoverage. Moreover, a comparison of the confidence intervals lengths clearly shows much more precise estimates for the EBP with augmentation compared to both other estimators. It should be noted that the accuracy of the SWEE confidence intervals is achieved at the expense of the longest average interval length. Interestingly, the coverage rate of the EBP without augmentation increases with the average confidence interval

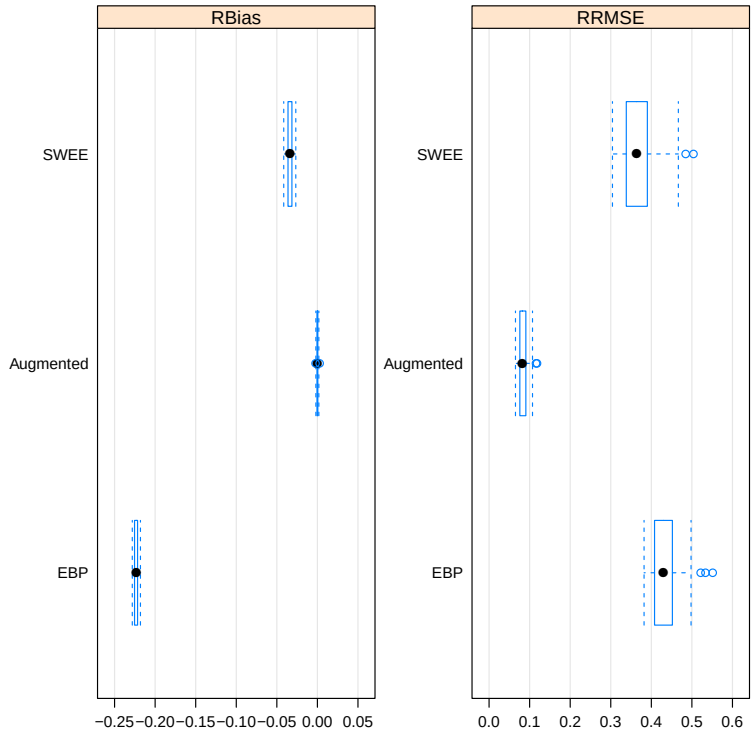

Figure 4.5: Quality of the point estimates under the PS size measure

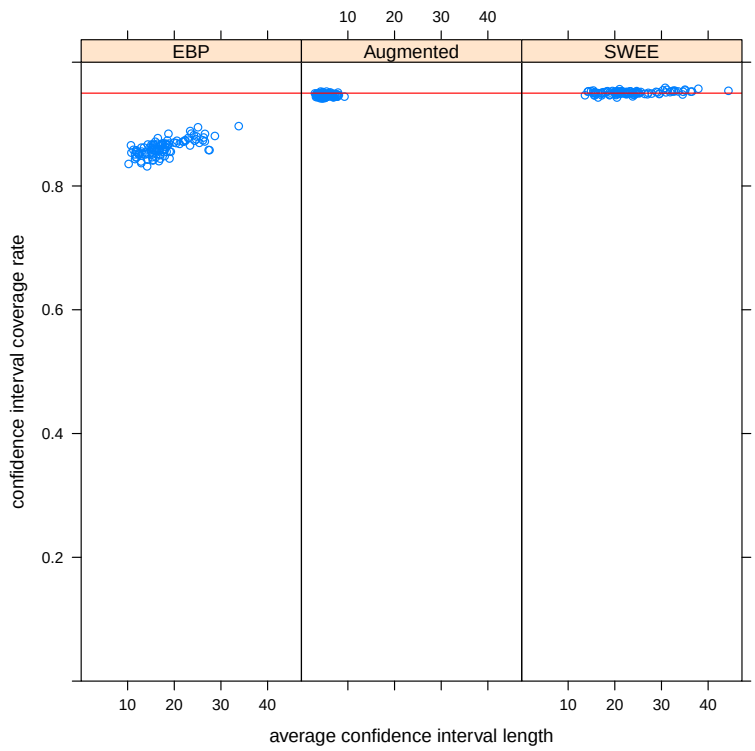

Figure 4.6: Confidence interval coverage rates under the PS size measure

length.

4.4.2 Asparouhov size measure

As a second case, we use size measure b_{dj} referred to as Asparouhov-type size measure by Verret et al. (2015) under invariant selection, as this strategy was inspired by the procedure of Asparouhov (2006). This yields the following size variable:

$$b_{dj} = \left[1 + \exp\left(-0.5\left\{\frac{1}{\alpha}\varepsilon_{dj} + \sqrt{1 - \frac{1}{\alpha^2}}\varepsilon^*_{dj}\right\}\right)\right]^{-1}. \tag{4.4.6}$$

In (4.4.6) $\varepsilon^*_{dj} \overset{i.i.d.}{\sim} N(0, \sigma^2_\varepsilon)$, $\mathrm{Cov}(\varepsilon_{dj}, \varepsilon^*_{dj}) = 0 \ \forall \ d, j$ and $\alpha \in \{1; 2; 3; \infty\}$. Note that α controls the degree of informativeness associated with the selection probabilities implied by (4.4.6). For $\alpha = 1$ the b_{dj} solely depend on the individual error term ε_{dj}, which gives rise to highly informative samples. On the other extreme, for $\alpha = \infty$, the selection probabilities do not depend on ε_{dj} at all, hence the sampling mechanism is ignorable.

We now illustrate the choice of the augmenting variable for the case of $\alpha = 1$. Scatterplots of the OLS residuals obtained from fitting the model (4.4.4) against different choices of the augmenting variable are shown in Figure 4.7. As the relationship between the design weights w_{dj} and the OLS residuals is clearly non-linear, they should not be used to augment the model. All other choices indicate a much better linear fit between the variables that can be used to augment the model and the residuals. While some non-linearities can be detected when plotting the residuals against the p^{-1}_{dj} and $\log(p_{dj})$, the scatterplot against the p_{dj} indicates a linear relationship.

We analyse the cAIC to assess the predictive accuracy of the different choices to augment the model. The results are tabulated in Table 4.4. It can be seen that using the p_{dj} clearly yields the highest predictive accuracy.

Table 4.4: cAIC for augmented modelling under the Asparouhov size measure with $\alpha = 1$

	Augmented variable				
	none	w_{dj}	p_{dj}	p^{-1}_{dj}	$\log(p_{dj})$
$\log g(\mathbf{y}\|\widehat{\boldsymbol{\xi}}(\mathbf{y}), \widehat{\mathbf{v}}(\mathbf{y}))$	-769.13	-25.09	1477.80	151.19	582.47
K	79.92	101.10	101.96	100.16	101.46
cAIC	1698.08	252.38	-2751.68	-102.07	-962.02

Finally, we have a look at the QQ plots of the transformed residuals to investigate the normality assumptions on v_d and ε_{dj} shown in Figure 4.8. It is easily seen that all of the choices violate the normality assumptions. Hence, we decided to focus on the p_{dj} as augmenting variable, because this choice clearly yields the highest predictive accuracy. Nonetheless, it violates the normality assumptions which may have an impact on the validity of point and MSE estimates.

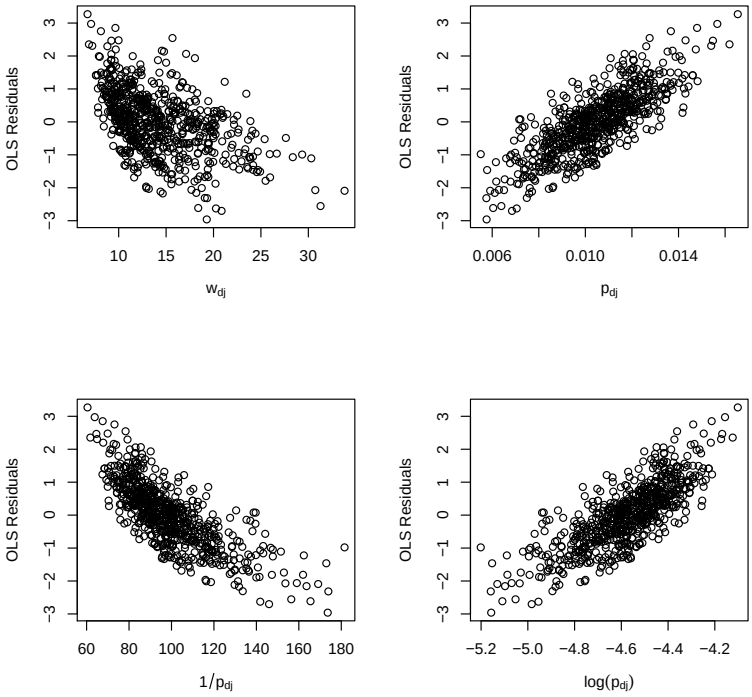

Figure 4.7: OLS residuals plotted against various choices of the augmenting variable under the Asparouhov size measure with $\alpha = 1$

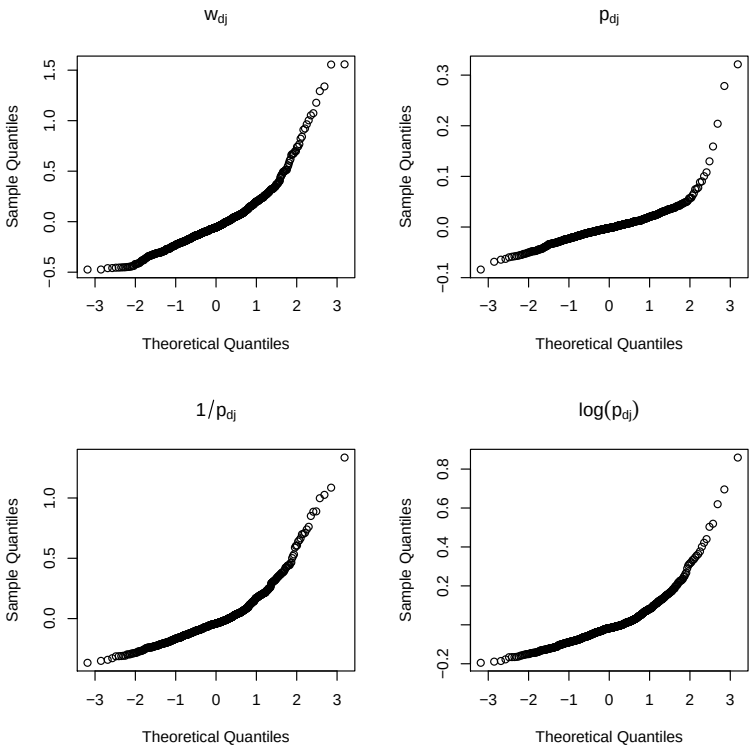

Figure 4.8: QQ plots of the transformed residuals for various choices of the augmenting variable under the Asparouhov size measure with $\alpha = 1$

For the sake of brevity the diagnostics to choose the augmenting variable for the cases of $\alpha = 2$ and $\alpha = 3$ are summarised in the Appendix B. Note that for both levels of informativeness, the differences in the predictive accuracy between augmenting with p_{dj}, p_{dj}^{-1} and $\log(p_{dj})$ are rather small with slight advantages for p_{dj}. Moreover, the normality assumption on the transformed residuals does not seem to be an issue for any of the choices. Hence, to facilitate a simple comparison we decided to use p_{dj} as augmenting variable for all values of α.

As the normality assumptions on the random part of the model seems questionable for $\alpha = 1$, we report the share of samples for which the Shapiro-Wilk test rejected the null hypothesis of normality at a significance level of 5% in Table 4.5. We observe that the normality assumption is rejected in every sample for the EBP under the augmented model when $\alpha = 1$. Thus, as already indicated by the QQ plot of the transformed residuals in Figure 4.8, including the selection probabilities among the model covariates has two implications. On the one hand, it increases the predictive accuracy of the model (cf. Table 4.4). On the other hand, the normality assumption, which could be maintained for the simple model with a rejection in only slightly over 5% of the samples, is no longer justifiable for the augmented model. Furthermore, note that the normality assumption for the augmented model does not seem to be a major concern for higher values of α. This agrees with the conclusions drawn from the QQ plots shown in the Appendix B.

Table 4.5: Share of samples under the Asparouhov size measure, in which the normality assumption on the transformed residuals was rejected on a significance level of 5%

α	EBP	Augmented
1	0.0512	1.0000
2	0.0508	0.0489
3	0.0488	0.0520
4	0.0498	0.0490

The relative biases of the estimates under the Asparouhov size measure are presented in Figure 4.9. We observe severe overestimation of the domain means using the EBP without augmentation for $\alpha = 1$. These biases reduce as the informativeness of the sampling mechanism decrease, i.e. as α increases. Finally, for the non-informative sampling mechanism the EBP yields unbiased estimates as expected. Using the SWEE estimator alleviates the biased of the EBP without augmentation to some extent. Nonetheless, the estimates are biased for all informative sampling mechanisms. Very interesting is the behaviour of the EBP under the augmented model. For the highly informative sampling mechanism with $\alpha = 1$ it yields very small but systematic negative biases. For all other values of α, however, systematic biases cannot be detected. Furthermore, it is seen that for $\alpha = \infty$ all estimation methods yield very similar and unbiased results.

A plausible explanation for this finding is that for the samples drawn using the highly informative sampling mechanism, the normality assumption on the transformed residuals under the augmented model could not be maintained.

The precision of the point estimates as measured by the RRMSE is shown in Figure 4.10. The most striking aspect is the very good performance of the EBP under the augmented model for $\alpha = 1$. In this setting, its RRMSEs are below 10 per cent in each domain. As the informativeness of the sampling mechanism decreases, the RRMSEs of the EBP under

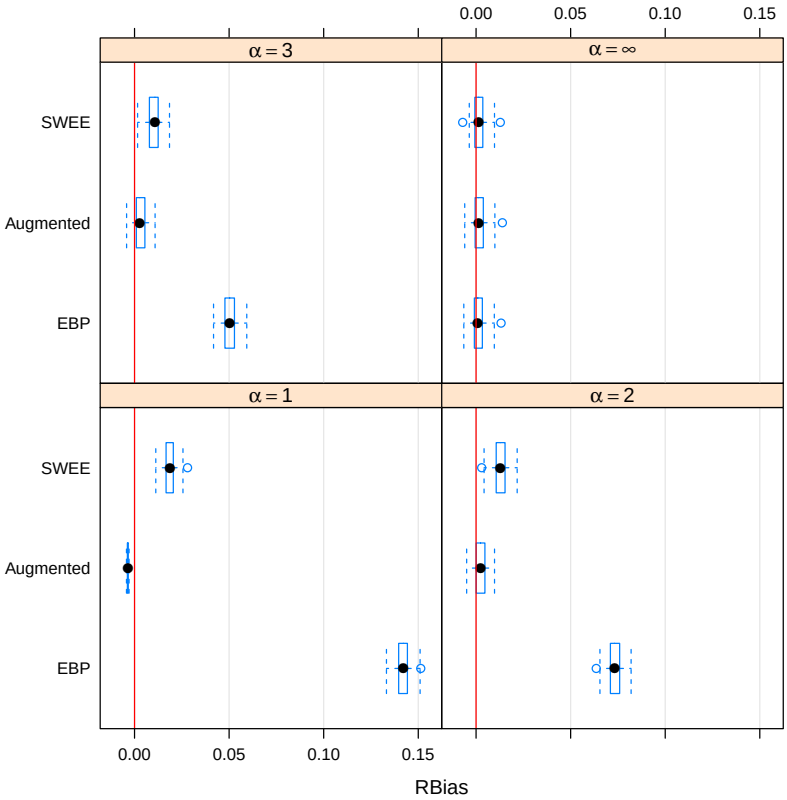

Figure 4.9: Relative biases under the Asparouhov size measure

the augmented model increase. This result stems from the fact that the variable used to augment the model underlying the EBP is highly predictive for the response variable when $\alpha = 1$. As α increases, the predictive power of the augmenting variable decreases and hence the EBP under the augmented model can improve less upon its non-augmented counterpart. Besides, slightly lower RRMSEs of the SWEE predictor compared to the EBP without augmentation can be observed under informative sampling mechanisms. The differences between these two estimation methods are not very pronounced in general.

In order to get a comprehensive picture of the quality of the confidence interval estimation, we study the coverage rates in connection with the average interval lengths. This is depicted in Figure 4.11. While there is undercoverage for the EBP under the augmented model when $\alpha = 1$, it is accompanied by much shorter confidence intervals lengths compared to the other estimation methods. Hence, a trade-off between precision and reliability of the confidence interval estimates occurs. Moreover, for the augmented EBP, we see that the average confidence interval lengths increase, as the degree of informativeness decreases. For other choices of α, the coverage rates of the EBP with augmentation meets the nominal rate. Furthermore, for $\alpha = 2$, which relates to a moderate level of informativeness, advantages of the augmented EBP due to smaller average interval lengths can be observed. Regarding the EBP without augmentation we may note a slight overcoverage for informative sampling mechanisms. For the case of a non-informative design the EBP without augmentation closely meets the desired coverage rates. Finally, using the SWEE estimator, very similar results are obtained for either value of α.

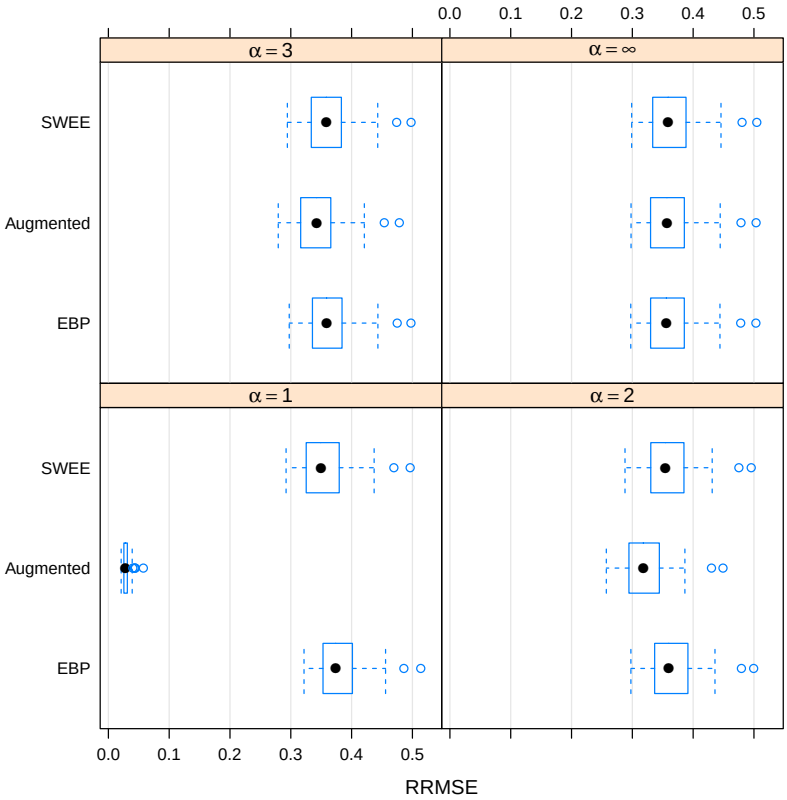

Figure 4.10: RRMSEs under the Asparouhov size measure

4.5 Summary and discussion

In this chapter, several strategies to incorporate the sampling design into model-based estimation procedures have been discussed. In Section 4.2, we reviewed approaches that are applicable to account for the design in the nested error regression model (3.2.1). Afterwards, in Section 4.3, we proposed two small area estimation methods to include the survey design into the log-transformed unit level model. The first of these approaches has been to augment the auxiliary information by a suitable function of the selection probabilities, which leads to the EBP under the augmented model. Our second proposal used survey-weighted estimating equations in the spirit of You and Rao (2002a) to estimate the vector of regression coefficients. Moreover, we provided a practical approach to select the augmenting variable based on a joint assessment of residual plots, a measure of predictive accuracy and a check of the normality assumptions.

The simulation studies revealed advantages of the EBP with an augmented model when compared to both the SWEE predictor and the simple EBP under informative sampling. The latter estimator was shown to suffer from severe systematic biases in the presence of informative designs. This finding was expected since the EBP without augmentation implicitly assumes that the sample and the population model coincide. Moreover, the SWEE predictor was shown to alleviate the biases inherent in the simple EBP to some extent. However, as soon as the RRMSEs of the two methods were considered, the advantages of the SWEE predictor over the non-augmented EBP diminished considerably. A potential cause of the rather limited gains of the SWEE predictor is that the studied

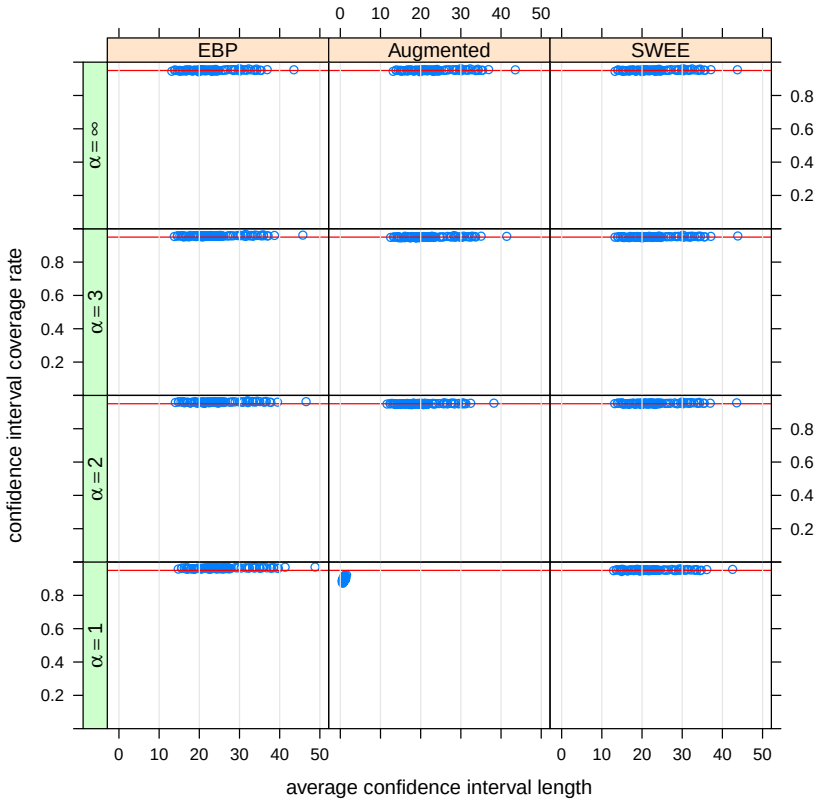

Figure 4.11: Confidence interval coverage rates under the Asparouhov size measure

sampling mechanisms were highly informative for functions of the selection probabilities p_{dj}, but to a lesser extent for the weights w_{dj}. Hence, using survey-weighted parameter estimates are not the best choice to correct for biases due to informative sampling in these settings.

In a similar vein, using the weights to augment the model was clearly dominated by other choices of the augmenting variable as indicated by model diagnostics and model selection tools. Under the PS size measure these tools identified $\log(p_{dj})$ as a suitable candidate to augment the model. The results of the simulation study supported this decision as the EBP under the augmented model provided unbiased point estimates with a high precision. Besides, also the confidence intervals reached the nominal coverage rates and at the same time, the average interval width was substantially smaller than for either of the other methods. Under the Asparouhov size measure, all options considered to augment the model suffered from a violation of the normality assumption in the highly informative setting with $\alpha = 1$. We chose p_{dj} to augment the model which led to very precise estimates in terms of the RRMSEs, but the estimates were systematically negatively biased. In addition to that the EBP under the augmented model did not meet the nominal coverage rates. Thus, the simulation results highlight the importance of checking the normality assumptions, as they can be violated under the augmented model even if they hold for the non-augmented model (cf. Table 4.5). Furthermore, it should be noted that a value of $\alpha = 1$ implies that the size variable and thus the selection probabilities only depend on the individual error term. Hence, augmenting the model by the selection probabilities will almost inevitably yield a much higher predictive accuracy compared to the non-augmented

model. A direct consequence of this is that the variance of the augmented estimates will be very small. As a side effect, however, the normality assumptions on the transformed residuals can no longer be maintained which leads to biases of the estimates, as the normality assumption is exploited in deriving the empirical best predictor, see (4.3.5).

Moreover, it should be noted that the augmented modelling approach could be extended to account for an informative sampling of areas as well. In this case, including the selection probabilities of the areas as well among the covariates can be a remedy.

Another very important issue is the coherence of small area estimates with aggregate estimates. While the pseudo-EBLUP (4.2.11) due to You and Rao (2002a) automatically satisfies the benchmarking property with respect to the national total, this does not hold for the estimators based on the lognormal unit level model presented in this chapter. A simple approach to fulfil this property is obtained by the following procedure:

$$\widehat{\mu}_d^{Bench} = \widehat{\mu}_d^{EBPLog} \frac{\widehat{\tau}^{Design}}{\widehat{\tau}^{EBPLog}}, \quad d = 1, \ldots, D, \tag{4.5.1}$$

where $\widehat{\tau}^{Design}$ denotes the estimated national total using a design-based or model-assisted estimator and $\widehat{\tau}^{EBPLog} = \sum_d N_d \widehat{\mu}_d^{EBPLog}$ (Berg & Chandra, 2014, p. 165). Obviously, the resulting benchmarked predictors $\widehat{\mu}_d^{Bench}$ are coherent on the national level with the estimated national total, which is an important issue for statistical agencies. In contrast to the EBP under the augmented model, knowing the design weights for the sampled units is sufficient, if a HT or GREG estimator of the total is used. Thereby, (4.5.1) is very simple to implement. It should be further noted that the benchmarking procedure can be applied on an intermediate level as well. In this case, the small area estimates are consistent with design-based estimates on an intermediate level. A drawback of the benchmarked predictor is that the estimation of its MSE is not straightforward. As noted by Berg and Chandra (2014) benchmarking can be applied to reduce the bias due to the non-linear back-transformation used in an MSE-optimal predictor as well. However, the bias reduction is achieved at the expense of an increase in the MSE relative to the EBP (Berg & Chandra, 2014, p. 165).

Chapter 5

Variance reduction with non-informative designs

In this chapter, we aim to create a sampling design which leads to precise design-based estimates for aggregations while not distorting the sampling model. A well-known strategy to reduce the variance in Monte-Carlo simulations is to use antithetic variables, where pairs of negatively correlated variables are drawn (cf. Rizzo, 2007, Section 5.4). We apply this principle to the context of survey sampling by drawing pairs of observations, where the pairs are constructed in an antithetic manner with respect to an auxiliary variable. In the terminology of Valliant et al. (2000), the resulting sample will aim at balance with respect to the auxiliary information. Compared to SRS, this may lead to a variance reduction for design-based estimators, provided that the auxiliary information is related to the dependent variable. The behaviour of the proposed strategy will be compared to other variance reducing designs, such as unequal probability sampling and stratification, in a number of simulation scenarios. Parts of this chapter have also been published in Zimmermann (2017). However, this chapter provides more detail on theoretical aspects of the proposed mechanism and uses other simulation studies.

5.1 Efficiency in a design-based context

For many of the designs discussed in Chapter 2, such as the optimal allocation under StrRS or πps sampling, it has been argued that they lead to a variance reduction compared to the benchmark case of SRS. While this may be true in many cases, it should be noted that statements of this kind are only possible if we consider an implicit model at hand (cf. Pfeffermann, 2011). To make this point clear, think about a survey planner aiming at variance reduction using design-based methods. Now suppose the survey he wants to conduct relates to an entirely new phenomenon about which little is known inside the agency and where also external knowledge is scarce. In this case, achieving a variance reduction versus SRS will be a very difficult task, as no prior model information is known. On the other hand, suppose that for some survey we have register information from previous years. We could easily use this knowledge to construct strata or possibly compute inclusion probabilities for an unequal probability scheme. This is likely to cause a variance reduction compared to the SRS case. It should be noted that the potential efficiency gains

are going to materialise, if and only if, the register information bears valuable knowledge about the target quantity. To put it differently, this method may work if the outcome can be implicitly viewed as a function of the register variables. If there is no such relationship, then design optimisation will not lead to a variance reduction. Moreover, in some cases, it may even lead to a loss in precision compared to SRS. An example in this regard is UPS, when the variable of interest is inversely proportional to the size variable. Hence, it is important to realise that efficiency can be achieved in a design-based context, provided the implied model holds.

Consider first the case of UPS and suppose that the size variable $z_k \in \mathbb{N}$ used to compute the inclusion probabilities via

$$\pi_k = n \cdot \frac{z_k}{\sum_{l \in U} z_l} = n \cdot \frac{z_k}{\tau_z} \tag{5.1.1}$$

is proportional to the variable of interest, i.e.

$$y_k \propto z_k.$$

In this case, the variance of the total using a Horvitz-Thompson estimator under a fixed-size design will be zero as pointed out by Särndal et al. (1992, p. 88), because

$$\begin{aligned}
\widehat{\tau}_y^{\mathrm{HT}} &= \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \frac{1}{n} \sum_{k=1}^{n} \frac{y_k \tau_z}{z_k} \\
&= \frac{\tau_z}{n} \sum_{k=1}^{n} \frac{y_k}{c y_k} \\
&= \frac{c \tau_y}{n} \frac{n}{c} = \tau_y,
\end{aligned}$$

regardless of the specific sample drawn. Hence, unequal probability sampling with πps will be the optimal choice, assuming an implicit model of the kind $y_k \propto z_k$, which can be rewritten as $y_k = c \cdot z_k$, where c is a constant.

Moreover, in the case of StrRS a variance reduction compared to SRS, i.e. a design effect smaller than, is frequently expected. It should be noted, however, that this holds only if the stratification is related to the dependent variable. In business surveys, for example the stratification is often based on a cross-classification of the company size in terms of the classified numbers of employees, the industry classification and some geographical information (cf. Hidiroglou & Lavallee, 2009). The reason why this is done is that the survey planner expects the variables of interest to vary with these stratification variables. This assumption may be plausible if the target variables comprise turnover, productivity and labour costs.

5.2 Antithetic clustering

5.2.1 Motivation

In the preceding section, we discussed efficiency from a design-based perspective. Thus, the reasoning applies to the estimation in planned domains with a large sample size.

Following the arguments in this work, model-based estimators are commonly applied in unplanned domains with small sample sizes. Hence, a mix of design- and model-based procedures for domains of different sizes is frequently adopted in practice (Little, 2012, p. 320). This raises the question of choosing the sampling design when the use of design-based as well as model-based estimation strategies can be anticipated ex ante. On the one hand, traditional procedures of design optimisation as presented in Chapter 2 may interfere with the models needed for small area estimation. This may cause serious problems such as informative sampling, which has been discussed in Chapter 4. On the other hand, simple ignorable designs such as SRS do not invalidate the model, but are prone to inefficient estimates on aggregate statistics. Hence, a compromise reflecting the trade-off between providing efficient design-based estimates for planned domains and enabling reliable model-based estimates for small areas is called for. Thus, our aim is to create a sampling design which can realise efficiency gains for design-based estimators while not distorting the properties of model-based procedures. Realising both goals simultaneously should permit reliable domain estimates using model-based methods and precise national estimates using design-based or model-assisted strategies. Optimal designs for small area estimation using a model-assisted approach are investigated by Molefe (2011), where the working model does not include auxiliary information. This is in contrast to our proposal, which explicitly allows for the correlation of the variable of interest with auxiliary information. Furthermore, our approach is related to the strategies discussed by Valliant et al. (2000, Section 3.4), who give a thought-provoking discussion of optimal sampling designs when the expectation is with respect to a model. They highlight the importance of selecting samples which are balanced for an auxiliary variable, as to reduce the conditional biases of design-based estimators. Our approach differs from the methods discussed by Valliant et al. (2000), as we explicitly account for domain estimation. Besides, our approach automatically yields ignorable sampling designs whenever the design on hierarchically higher levels is ignorable as well.

From now on, we assume that a size variable z_k exists which has to be known for all $k = 1, \ldots, N$ units in the population and can be obtained from a register, or the last census. Furthermore, we assume that the order statistic can be produced for z_k, i.e. we can rank z_k in increasing order. Now the idea is to cluster our variable in clusters of size 2, such that there are $L = \lceil N/2 \rceil$ clusters. To increase the efficiency of a design-based method, we cluster the elements in an antithetic fashion. This leads composition of the clusters detailed in Definition 1

Definition 1. *Suppose the primary sampling units (PSUs) are constructed such that l_r denotes the r-th PSU, which comprises*

$$l_1 = \{z_{[1]}, z_{[N]}\};\ l_2 = \{z_{[2]}, z_{[N-1]}\}\ \ldots; l_L = \{z_{[N/2]}, z_{[N/2+1]}\}; \tag{5.2.1}$$

for even N and

$$l_1 = \{z_{[1]}, z_{[N]}\};\ l_2 = \{z_{[2]}, z_{[N-1]}\}\ \ldots; l_L = \{z_{[\lceil N/2 \rceil]}\}; \tag{5.2.2}$$

for odd N, where $z_{[k]}$ refers to the $k-$th element of the order statistic. A construction scheme satisfying (5.2.1) or (5.2.2) will be called antithetic clustering (ATC).

Note that our approach leads to PSUs of size two if N is even. If N is odd, $L - 1$ PSUs are of size two, while the PSU comprising the median of z_k is of size one. While Definition 1 specifies how to construct the PSUs, the question of how to draw the sample of PSUs

is still left open. Our proposal is to gather the sample by means of a SRS of clusters, where l out of L PSUs are drawn. Hence, the resulting sample size in terms of ultimate sampling units included in the sample is

$$n^{ATC} = \begin{cases} 2l, & \text{if N is even,} \\ 2l-1, & \text{else.} \end{cases} \tag{5.2.3}$$

Note that the idea to base the sample design on the sorted values of an auxiliary variable is not new. A systematic sampling approach based on the order statistic $z_{[k]}$ is discussed in Valliant et al. (2000, Section 3.4.2) and references therein. Their proposal for systematic sampling with equal probabilities is outlined in Algorithm 3.

Algorithm 3 Systematic sampling with equal probabilities

1. Sort the elements according to the auxiliary information z_k.
2. Compute the sampling interval a as N/n.
3. Draw randomly an integer $r \in [1; a]$ with equal probabilities.
4. The sample consists of the elements k from the ordered register satisfying $k = r + (i-1)a$ with $k \leq N$, $i = 1, \ldots, n$.

Note that the design detailed in Algorithm 3 is a special case of a SIC design where only one cluster is selected. Thus, unbiased variance estimation under a design-based approach is infeasible with this design as the second-order inclusion probabilities $\pi_{kl} = 0$ for elements which do not belong to the same cluster (Särndal et al., 1992, p. 75). Instead, one option is to resort to model-based variance estimators, which can be robustified against model misspecification by non-parametric estimation (cf. Opsomer, Francisco-Fernández, & Li, 2012). This limitation does not occur with our approach, as we draw a SRS of clusters, where more than one cluster is sampled. Nonetheless, systematic sampling with equal probabilities can be suitable for model-based estimation as it fulfils (4.1.2). As we want to estimate national statistics with a design-based or model-assisted approach, unbiased variance estimates are clearly desired. Hence, the approach of Algorithm 3 is not sufficient for our purposes.

Obviously, in our approach all PSUs and hence all units k have the same probability of being included in the sample. Since the sample selection is independent of the y_k values, this design is ignorable by (4.1.2). The question that remains is whether this sampling mechanism is suitable for design-based estimation. To analyse the consequences, we study the properties of the sample mean under cluster sampling.

5.2.2 Suitability of single stage cluster sampling

In the following, we focus on the variance of the sample mean under SRS and SIC with an SRS of PSUs. We will see how these expressions are related by the intraclass correlation coefficient and review conditions under which SIC can lead to efficiency gains compared

to SRS. To ease the notation, we will discuss the case of estimating the national mean first, before we move on to the estimation of domain means.

In the following, we establish a condition under which a SIC design with a SRS of equally sized PSUs, each of size $\overline{N}_L$, has a smaller variance for the sample mean than a SRS scheme. Intuitively, this requires elements within PSUs to vary more than elements between clusters. A commonly used measure for the homogeneity within clusters is the intraclass correlation coefficient (ICC), which can be used to derive the design effect under SIC. The ICC is given by (Lohr, 1999, Section 5.2.2):

$$\mathrm{ICC} = 1 - \frac{\overline{N}_L}{\overline{N}_L - 1}\frac{\mathrm{SSW}}{\mathrm{SSTOT}}, \tag{5.2.4}$$

where SSTOT denotes the total sum of squares, i.e

$$\mathrm{SSTOT} = \sum_{k=1}^{N}(y_k - \overline{Y})^2,$$

with $\overline{Y} = N^{-1}\sum_{k=1}^{N} y_k$ as the overall mean of the variable of interest. Further, SSW denotes the sum of squares within the clusters given by

$$\mathrm{SSW} = \sum_{h=1}^{L}\sum_{k=1}^{N_h}(y_{hk} - \overline{Y}_h)^2,$$

with N_h as the size of the h-th PSU and $\overline{Y}_h$ as the mean of the variable of interest in cluster h. Since the total sum of squares can be partitioned into the SSW and the sum of squares between clusters (SSB), where

$$\mathrm{SSB} = \sum_{h=1}^{L} N_h(\overline{Y}_h - \overline{Y})^2,$$

it follows that $-\frac{1}{\overline{N}_L - 1} \leq \mathrm{ICC} \leq 1$. The design effect (2.3.1) under SIC with SRS of equally sized PSUs follows as (Lohr, 1999, Section 5.2.2):

$$\mathrm{DEFF}_{\mathrm{SIC}} = \frac{\mathrm{Var}(\widehat{\mu}_{\mathrm{SIC}})}{\mathrm{Var}(\widehat{\mu}_{\mathrm{SRS}})} = \frac{\overline{N}_L L - 1}{\overline{N}_L(L-1)}\left(1 + (\overline{N}_L - 1)\mathrm{ICC}\right). \tag{5.2.5}$$

SIC is more efficient than SRS if $\mathrm{DEFF}_{\mathrm{SIC}} < 1$. It can be easily seen that this translates to the following condition:

$$\mathrm{ICC} < \frac{-1}{L\overline{N}_L - 1}. \tag{5.2.6}$$

It follows from (5.2.6) that for SIC to be more precise than SRS, we need a negative ICC. To obtain an expression in terms of the sum of squares for the case that SIC is more

precise than SRS, we combine (5.2.4) with (5.2.6) to get

$$\begin{aligned}
&\text{ICC} && < \frac{-1}{\overline{N}_L L - 1} \\
&\Rightarrow 1 - \frac{\overline{N}_L}{\overline{N}_L - 1}\frac{\text{SSW}}{\text{SSW}+\text{SSB}} && < \frac{-1}{\overline{N}_L L - 1} \\
&\Rightarrow \frac{\overline{N}_L}{\overline{N}_L - 1}\frac{\text{SSW}}{\text{SSW}+\text{SSB}} && > \frac{\overline{N}_L L}{\overline{N}_L L - 1} \\
&\Rightarrow \text{SSW} && > \frac{L(\overline{N}_L - 1)}{\overline{N}_L L - 1}(\text{SSW}+\text{SSB}) \\
&\Rightarrow \text{SSW}\left(1 - \frac{L(\overline{N}_L - 1)}{\overline{N}_L L - 1}\right) && > \frac{L(\overline{N}_L - 1)}{\overline{N}_L L - 1}\text{SSB} \\
&\Rightarrow \text{SSW} && > \frac{L(\overline{N}_L - 1)}{L - 1}\text{SSB}.
\end{aligned} \tag{5.2.7}$$

Hence the sum of squares within has to be greater than the sum of squares between times a factor which depends on the number of PSUs and the size of the PSUs. For the special case of $\overline{N}_L = 2$, which occurs under ATC with an even number of elements in the population, this constant reduces to $L/(L-1)$. Furthermore, we see that as the cluster size $\overline{N}_L$ increases, the ratio of the within cluster variation relative to the between cluster variation has to rise as well, if SIC is to be more efficient than SRS. Hence, our proposal to construct clusters of size two is the least demanding for SIC to yield estimates with a greater precision than SRS. An implication of (5.2.7) is to create clusters such that most of the variation of the dependent variable is due to variation within the clusters, not between clusters. What does this imply for our antithetic construction of PSUs based on the auxiliary variable z_k? Intuitively, the PSUs defined by (5.2.1) will have a large ratio of the within versus the between variation for the auxiliary variable.

Proposition 1. *Suppose the number of elements in the population is even and that the clusters are formed such that one unit with a below-median value is coupled with one unit of an above-median value. If these conditions are met, the ATC approach minimises the SSB of the clustering variable among all possible combinations of PSUs that are exhaustive and mutually exclusive.*

Proof. Observe that the clustering is done with respect to the order statistic $z_{[k]}$. For notational convenience, we denote the vector of elements with a below-median value of z_k by the L-tuple $\mathbf{a}$, which is ordered, such that $a_{[1]} \leq \ldots \leq a_{[L]}$. Similarly, the elements with an above-median value of z_k belong to the L-tuple $\mathbf{b}$, such that $b_{[1]} \leq \ldots b_{[L]}$. Minimising the SSB under these conditions is equivalent to

$$\min F = \sum_{h=1}^{L}(\overline{Z}_h - \overline{Z})^2 = \sum_{h=1}^{L}\left(\frac{1}{2}(a_h + b_h) - \overline{Z}\right)^2,$$

where $\overline{Z}_h$ is the mean of the z_k belonging to cluster h and $\overline{Z}$ denotes the population mean of the z_k. Consider any deviation from the ATC scheme in the vector $\mathbf{b}$, which implies a permutation, i.e. $b_{[v[h]]}$ and denote this allocation as PER. The loss functions can be rewritten as

$$F_{\text{ATC}} = \sum_{h=1}^{L} \left(\frac{1}{2} \left(a_{[h]} + b_{[L-h+1]} \right) - \overline{Z} \right)^2 = \sum_{h=1}^{L} \frac{1}{4} \left(a_{[h]} + b_{[L-h+1]} \right)^2 - L\overline{Z}^2 \tag{5.2.8}$$

$$F_{\text{PER}} = \sum_{h=1}^{L} \left(\frac{1}{2} \left(a_{[h]} + b_{[L-\upsilon(h)+1]} \right) - \overline{Z} \right)^2 = \sum_{h=1}^{L} \frac{1}{4} \left(a_{[h]} + b_{[L-\upsilon(h)+1]} \right)^2 - L\overline{Z}^2. \tag{5.2.9}$$

Hence, the condition under which ATC is optimal is given by

$$\sum_{h=1}^{L} \left(a_{[h]} + b_{[L-h+1]} \right)^2 \leq \sum_{h=1}^{L} \left(a_{[h]} + b_{[L-\upsilon(h)+1]} \right)^2 \tag{5.2.10}$$

which due to the requirement of an exhaustive allocation reduces to

$$\sum_{h=1}^{L} a_{[h]} b_{[L-h+1]} \leq \sum_{h=1}^{L} a_{[h]} b_{[L-\upsilon(h)+1]}. \tag{5.2.11}$$

We know from the rearrangement inequality (cf. Hardy, Littlewood, & Pólya, 1952, p. 261) that the inner product of two vectors **a** and **b** attains its lower bound when the vector **a** is sorted in increasing order and the vector **b** in decreasing order. This is the case with the ATC scheme. Hence, the RHS of (5.2.11) will be equal to the LHS iff the permutation is the identity, i.e. $\upsilon[h] = h$ and larger otherwise. □

5.2.3 Antithetic clustering under single level models

Having established a certain optimality of ATC in the previous section, we need to examine the consequences for our variable of interest. In the case that the y_k are a linear function of the z_k, i.e. they are perfectly correlated, ATC yields better results than SRS, whenever condition (5.2.7) is satisfied for the clustering variable. Since this assumption is hardly sustainable in practice, we have to consider models specifying the data generating process.

Suppose the relationship between the dependent variable and the auxiliary information used for clustering is given by a linear regression model, i.e.

$$\begin{aligned} y_k &= \beta_0 + \beta_1 z_k + \varepsilon_k, \\ \varepsilon_k &\overset{i.i.d}{\sim} G(0, \sigma^2) \end{aligned} \tag{5.2.12}$$

where ε_k denotes the error term, which is assumed to be independently and identically distributed according to a distribution G with mean zero and variance σ^2. Amongst the choices for G are the normal distribution, different t-distributions etc.

To derive the expectation of SSW and SSB under the model (5.2.12), it is useful to simplify the two expressions. The sum of squares between for the dependent variable is given by

$$\text{SSB}_Y = \sum_{h=1}^{L} N_h (\overline{Y}_h - \overline{Y})^2 = \sum_{h=1}^{L} N_h \overline{Y}_h^2 - N\overline{Y}^2 \tag{5.2.13}$$

while the sum of squares within is defined as

$$\mathrm{SSW}_Y = \sum_{h=1}^{L}\sum_{k=1}^{N_h}(y_k - \overline{Y}_h)^2 = \sum_{h=1}^{L}\sum_{k=1}^{N_h} y_k^2 - \sum_{h=1}^{L} N_h \overline{Y}_h^2. \tag{5.2.14}$$

Now, the expectation of the SSB and SSW under model (5.2.12) can be calculated as

$$\begin{aligned} \mathrm{E}_M\left(\mathrm{SSB}_Y\right) &= \sum_{h=1}^{L} N_h \mathrm{E}_M\left(\overline{Y}_h^2\right) - N\mathrm{E}_M\left(\overline{Y}^2\right) \\ \mathrm{E}_M\left(\mathrm{SSW}_Y\right) &= \sum_{h=1}^{L}\sum_{k=1}^{N_h} \mathrm{E}_M\left(y_k^2\right) - \sum_{h=1}^{L} N_h \mathrm{E}_M\left(\overline{Y}_h^2\right), \end{aligned} \tag{5.2.15}$$

where $E_M(\cdot)$ denotes the expectation with respect to the model (5.2.12). It can be seen from (5.2.15) that we need expressions for the expected values of $\overline{Y}_h^2$, $\overline{Y}^2$ and y_k^2 under model (5.2.12). To compute these quantities, we employ the variance identity $\mathrm{E}(X^2) = [\mathrm{E}(X)]^2 + \mathrm{Var}(X)$. Applying this formula yields the following formulae

$$\begin{aligned} \mathrm{E}_M\left(y_k^2\right) &= \beta_0^2 + 2\beta_0\beta_1 z_k + \beta_1^2 z_k^2 + \sigma^2 \\ \mathrm{E}_M\left(\overline{Y}_h^2\right) &= \beta_0^2 + 2\beta_0\beta_1 \overline{Z}_h + \beta_1^2 \overline{Z}_h^2 + \sigma^2/N_h \\ \mathrm{E}_M\left(\overline{Y}^2\right) &= \beta_0^2 + 2\beta_0\beta_1 \overline{Z} + \beta_1^2 \overline{Z}^2 + \sigma^2/N. \end{aligned} \tag{5.2.16}$$

Inserting the second and third equation of (5.2.16) into the first equation of (5.2.15) yields:

$$\begin{aligned} \mathrm{E}_M\left(\mathrm{SSB}_Y\right) &= \sum_{h=1}^{L} N_h\left(\beta_0^2 + 2\beta_0\beta_1\overline{Z}_h + \beta_1^2\overline{Z}_h^2 + \sigma^2/N_h\right) - N\left(\beta_0^2 + 2\beta_0\beta_1\overline{Z} + \beta_1^2\overline{Z}^2 + \sigma^2/N\right) \\ &= \beta_1^2 \underbrace{\left(\sum_{h=1}^{L} N_h\overline{Z}_h^2 - N\overline{Z}^2\right)}_{\mathrm{SSB}_Z} + \sigma^2\left(L-1\right) \end{aligned} \tag{5.2.17}$$

Analogously, we can combine the first two equations of (5.2.16) with the SSW equation of (5.2.15) to obtain the expectation of the sum of squares within under model (5.2.12). This leads to

$$\begin{aligned} \mathrm{E}_M\left(\mathrm{SSW}_Y\right) &= \sum_{h=1}^{L}\sum_{k=1}^{N_h}\left(\beta_0^2 + 2\beta_0\beta_1 z_k + \beta_1^2 z_k^2 + \sigma^2\right) - \sum_{h=1}^{L} N_h\left(\beta_0^2 + 2\beta_0\beta_1\overline{Z}_h + \beta_1^2\overline{Z}_h^2 + \sigma^2/N_h\right) \\ &= \beta_1^2 \underbrace{\left(\sum_{h=1}^{L}\sum_{k=1}^{N_h} z_k^2 - \sum_{h=1}^{L} N_h\overline{Z}_h^2\right)}_{\mathrm{SSW}_Z} + \sigma^2\left(N-L\right). \end{aligned} \tag{5.2.18}$$

We assume now that N is even such that $L = N/2$, which implies $N - L = L$. Inserting expressions (5.2.18) and (5.2.17) into condition (5.2.7) yields:

$$\frac{\mathrm{E}_M(\mathrm{SSW}_Y)}{\mathrm{E}_M(\mathrm{SSB}_Y)} = \frac{\beta_1^2\mathrm{SSW}_Z + L\,\sigma^2}{\beta_1^2\mathrm{SSB}_Z + (L-1)\,\sigma^2} > \frac{L}{L-1}, \tag{5.2.19}$$

which assuming a non-zero value of β_1 can be simplified to

$$\frac{\mathrm{E}_M(\mathrm{SSW}_Y)}{\mathrm{E}_M(\mathrm{SSB}_Y)} = \frac{\mathrm{SSW}_Z}{\mathrm{SSB}_Z} > \frac{L}{L-1}. \tag{5.2.20}$$

Hence, ATC performs better than SRS under a linear model in expectation provided the correlation between the variable of interest is non-zero and the ratio of the within to the between variation in the auxiliary variable is greater than $L/(L-1)$, which approaches 1 for a large number of clusters L. If $\beta_1 = 0$, the variable of interest and the clustering variable are uncorrelated. It can be shown that in this case under model (5.2.12) ATC would be as precise as SRS.

Now, as a second case, we consider a slight modification of (5.2.12), where we introduce heteroscedastic error terms. The model reads:

$$\begin{aligned} y_k &= \beta_0 + \beta_1 z_k + \varepsilon_k, \\ \varepsilon_k &\overset{ind}{\sim} G(0, \sigma_k^2). \end{aligned} \tag{5.2.21}$$

We derive expressions for the expected values of $\overline{Y}_h^2$, $\overline{Y}^2$ and y_k^2 under model (5.2.21). They are given by

$$\begin{aligned} \mathrm{E}_M\left(y_k^2\right) &= \beta_0^2 + 2\beta_0\beta_1 z_k + \beta_1^2 z_k^2 + \sigma_k^2 \\ \mathrm{E}_M\left(\overline{Y}_h^2\right) &= \beta_0^2 + 2\beta_0\beta_1 \overline{Z}_h + \beta_1^2 \overline{Z}_h^2 + \frac{1}{N_h^2}\sum_{k=1}^{N_h} \sigma_k^2 \\ \mathrm{E}_M\left(\overline{Y}^2\right) &= \beta_0^2 + 2\beta_0\beta_1 \overline{Z} + \beta_1^2 \overline{Z}^2 + \frac{1}{N^2}\sum_{k=1}^{N} \sigma_k^2. \end{aligned} \tag{5.2.22}$$

This yields the following expressions for the expected sum of squares between:

$$\mathrm{E}_M\left(\mathrm{SSB}_Y\right) = \beta_1^2 \mathrm{SSB}_Z + \sum_{h=1}^{L} \frac{1}{N_h} \sum_{k=1}^{N_h} \sigma_k^2 - \frac{1}{N}\sum_{k=1}^{N} \sigma_k^2, \tag{5.2.23}$$

which in the case of equally sized PSUs, i.e. $N_h = \overline{N}_L \; \forall \; h$ reduces to

$$\mathrm{E}_M\left(\mathrm{SSB}_Y\right) = \beta_1^2 \mathrm{SSB}_Z + \left(\frac{1}{\overline{N}_L} - \frac{1}{N}\right) \sum_{k=1}^{N} \sigma_k^2.$$

The expected sum of squares within is given by

$$\mathrm{E}_M\left(\mathrm{SSW}_Y\right) = \beta_1^2 \mathrm{SSW}_Z + \sum_{k=1}^{N} \sigma_k^2 - \sum_{h=1}^{L} \frac{1}{N_h} \sum_{k=1}^{N_h} \sigma_k^2, \tag{5.2.24}$$

which in the case of equally sized PSUs reduces to

$$\mathrm{E}_M\left(\mathrm{SSW}_Y\right) = \beta_1^2 \mathrm{SSW}_Z + \left(1 - \frac{1}{\overline{N}_L}\right) \sum_{k=1}^{N} \sigma_k^2.$$

Assuming that N is even and all PSUs are of the same size, $N_h = \overline{N}_L = 2\ \forall h$, we can plug expressions (5.2.23) and (5.2.24) into condition (5.2.7) to get:

$$\frac{\mathrm{E}_M(\mathrm{SSW}_Y)}{\mathrm{E}_M(\mathrm{SSB}_Y)} = \frac{\beta_1^2 \mathrm{SSW}_Z + \frac{1}{2}\sum_{k=1}^N \sigma_k^2}{\beta_1^2 \mathrm{SSB}_Z + \frac{1}{2}\frac{L-1}{L}\sum_{k=1}^N \sigma_k^2} > \frac{L}{L-1}, \tag{5.2.25}$$

which assuming a non-zero value of β_1 can be simplified to

$$\frac{\mathrm{E}_M(\mathrm{SSW}_Y)}{\mathrm{E}_M(\mathrm{SSB}_Y)} = \frac{\mathrm{SSW}_Z}{\mathrm{SSB}_Z} > \frac{L}{L-1}. \tag{5.2.26}$$

Hence, assuming even N implies that the efficiency conditions under models (5.2.12) and (5.2.21) are identical.

5.2.4 Antithetic clustering under a model with domain effects

While the developments from the previous sections are based on a simple single level linear regression model, the condition (5.2.19) applies as well to a model with domain-specific effects v_d (which can be random or fixed)

$$\begin{aligned} y_k &= \beta_0 + \beta_1 z_k + v_d + \varepsilon_k, \quad k \in U_d \\ \varepsilon_k &\overset{i.i.d.}{\sim} G(0, \sigma^2) \end{aligned} \tag{5.2.27}$$

provided that the sampling design is a two-stage design with the domains as strata on the first stage (planned domains) and within domains the ATC procedure is applied. The reason why this holds is that within a domain d, the domain-specific effect v_d in (5.2.27) is a constant and thus absorbed by the intercept term. Thus, within domains, model (5.2.27) reduces to (5.2.21) with different intercepts. Since the national mean is a convex combination of the stratum means under StrRS, applying ATC within domains will lead to a better estimate for the national mean than SRS within domains, provided (5.2.25) holds.

Now suppose that the model governing the population is indeed given by (5.2.27), but ATC is applied on the population level directly. This leads to changes for the relevant expectations needed to compute the sum of squares between and within as the cluster can be composed of units from different domains. The relevant expectations now follow as:

$$\begin{aligned} \mathrm{E}_M\left(y_k^2\right) &= \beta_0^2 + 2\beta_0\beta_1 z_k + 2\beta_0 v_k + \beta_1^2 z_k^2 + 2\beta_1 z_k v_k + v_k^2 + \sigma^2 \\ \mathrm{E}_M\left(\overline{Y}_h^2\right) &= \beta_0^2 + 2\beta_0\beta_1 \overline{Z}_h + 2\beta_0 \overline{V}_h + \beta_1^2 \overline{Z}_h^2 + 2\beta_1 \overline{Z}_h \overline{V}_h + \overline{V}_h^2 + \sigma^2/N_h \\ \mathrm{E}_M\left(\overline{Y}^2\right) &= \beta_0^2 + 2\beta_0\beta_1 \overline{Z} + 2\beta_0 \overline{V} + \beta_1^2 \overline{Z}^2 + 2\beta_1 \overline{ZV} + \overline{V}^2 + \sigma^2/N, \end{aligned} \tag{5.2.28}$$

where v_k denotes the domain-specific effect relevant for unit k and $\overline{V}$ and $\overline{V}_h$ refer to the population and cluster means of the domain-specific effects, respectively. Using (5.2.28)

in connection with expressions for $\mathrm{E}_M(\mathrm{SSB}_Y)$ and $\mathrm{E}_M(\mathrm{SSW}_Y)$ yields:

$$\begin{aligned}
\mathrm{E}_M(\mathrm{SSB}_Y) &\approx \beta_1^2 \mathrm{SSB}_Z + (L-1)\cdot\sigma^2 + \underbrace{\sum_{h=1}^{L} N_h \overline{V}_h^2 - N\cdot\overline{V}^2}_{:=\mathrm{SSB}_V} \\
E_M(\mathrm{SSW}_Y) &\approx \beta_1^2 \mathrm{SSW}_z + L\cdot\sigma^2 + \underbrace{\sum_{d=1}^{D} N_d v_d^2 - \sum_{h=1}^{L} N_h \overline{V}_h^2}_{:=\mathrm{SSW}_V}.
\end{aligned} \tag{5.2.29}$$

Note that expressions (5.2.29) are approximations, since cross-product terms between the domain-specific effects and the clustering variable as well as those between the domain specific effects and the individual error terms are ignored. These approximations can be motivated by the fact that in many applications the cross-product terms are negligible compared to the terms present in (5.2.29). Exact formulae for the expected sum of squares within and between are given in Appendix C.1. From (5.2.29) and (5.2.7) the requirement for ATC to be more precise than SRS reads

$$\frac{\mathrm{E}_M(\mathrm{SSW}_Y)}{\mathrm{E}_M(\mathrm{SSB}_Y)} = \frac{\beta_1^2 \mathrm{SSW}_Z + \mathrm{SSW}_V}{\beta_1^2 \mathrm{SSB}_Z + \mathrm{SSB}_V} > \frac{L}{L-1}. \tag{5.2.30}$$

Hence, the domain specific effects v_d play a similar role to the clustering variable z_k. As a consequence, condition (5.2.30) will be harder to fulfil compared to (5.2.19). Note that to obtain expressions (5.2.28) we effectively treated the domain effects v_d as fixed. This can be achieved using a random effects specification as well, provided the expectations are taken conditional on the vector of domain effects. Moreover, we could calculate unconditional expectations under a random effect specification, but they are of little interest in practice.

5.3 Simulation studies

In this section, we report the findings from a quasi-design-based study which was conducted to enable a comparison between ATC and other sampling designs. This approach was chosen as it allows to isolate the impact of the sampling mechanism from potentially confounding effects such as model misspecification.

The finite population consists of $N = 12000$ elements which belong to $D = 30$ domains. To introduce variation in the domain sizes, we proceeded in a similar fashion to Lehtonen and Veijanen (2009, Section 5.2) and allocated the units to domains with probabilities proportional to $\exp(q_d)$ where $q_d \overset{ind}{\sim} U[0, 2.9]$. This led to the distribution of domain sizes as given in Table 5.1.

The population was generated according to the random intercept model

$$\begin{aligned}
&y_{dj} = 6 + 3\cdot x_{dj} + v_d + \varepsilon_{dj}, \quad j = 1,\ldots,N_d, \quad d = 1,\ldots,D, \\
&v_d \overset{i.i.d}{\sim} N(0,\sigma_v^2) \\
&\varepsilon_{dj} \overset{i.i.d}{\sim} N(0,\sigma_\varepsilon^2).
\end{aligned} \tag{5.3.1}$$

Table 5.1: Domain sizes N_d

Domain	N_d	Domain	N_d	Domain	N_d
01	57	11	212	21	474
02	74	12	206	22	559
03	62	13	229	23	646
04	103	14	256	24	711
05	124	15	260	25	724
06	123	16	238	26	830
07	114	17	246	27	946
08	113	18	356	28	968
09	181	19	387	29	1097
10	168	20	369	30	1167

For the shrinkage effect, it should be noted that since

$$\gamma_d = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\varepsilon^2/n_d} = \frac{1}{1 + \omega/n_d},$$

with $\omega \equiv \sigma_\varepsilon^2/\sigma_v^2$, γ_d depends only on the ratio of the variance components, but not the components itself. Hence, we set $\sigma_\varepsilon^2 = 4$ throughout the simulation and chose σ_v^2 to obtain $\omega = \{2, 4, 8\}$. These are values that are common in practice. The covariate x_{dj} was drawn as $x_{dj} \overset{i.i.d}{\sim} N(1, 1)$.

As estimators for the small area means we compare the model-based BHF predictor (3.2.8), a direct estimator and an indirect GREG estimator. In general, we use a fully combined fixed-effects assisting model for the GREG estimator, which does not account for differences beyond the covariates, since with planned domains the domain effects essentially cancel out as explained in Section 2.2. For the unplanned domains, however, we also consider a MGREG, where the predictions are obtained from the unit level random intercept model. Additionally, we also compared national estimates using the Direct and GREG (MGREG) estimators. The sampling designs which were considered, comprise ATC, SRS, StrRS and UPS plans with a total (expected) sample size of $n = 500$ units. As auxiliary information at the design stage, we take x_{dj}, the covariate used in specifying the model.

Besides the predictive power of the model and the shrinkage factor, we also analyse the impact of planned versus unplanned domain structures, i.e., are the sampling designs implemented on national or domain level.

5.3.1 Mean estimates under planned domains

In the first application, we investigate the use of the above mentioned designs, when the domains are planned at the design stage with known (expected) sample sizes. This implies that the sampling schemes treat the domains as strata and leads to two-stage designs. We consider three types of domain-specific sample sizes: the (expected) sample size in domains $d = 1, \ldots, 10$ is $n_d = 6$, for domains $d = 11, \ldots, 20$ we have $n_d = 14$ and for domains $d = 21, \ldots, 30$ we set $n_d = 30$. This implies values for γ_d as tabulated in Table

5.2. It can be seen that for the case with $n_d = 30$ and $\omega = 2$ the BHF predictor attaches most weight to the survey regression component, whereas for $n_d = 6$ and $\omega = 8$ more weight is given towards to the regression synthetic component. For the ATC design, we

Table 5.2: Implied values of γ_d in a quasi-design-based simulation study with planned domains

	$\omega = 2$	$\omega = 4$	$\omega = 8$
$n_d = 6$	0.75	0.60	0.429
$n_d = 14$	0.875	0.778	0.636
$n_d = 30$	0.938	0.882	0.789

should note that in case of even domain sizes, the sample sizes will be equal to n_d, whereas with odd domain sizes, a sample size can take a value of $n_d - 1$ if the cluster that contains only the median is drawn. In the StrRS design, we construct strata within domains along the ordered values of x_{dj}. In domains with a sample size of 6, we construct two strata from which 3 elements are drawn. In domains with a sample size of 14, four strata are constructed, two of them having a sample size of 3 and the other two a sample size of 4 units per stratum. Finally, domains with $n_d = 30$ are subdivided into 5 strata with sample sizes of 6 units per stratum. The stratum-specific population sizes $N_{d,h}$ were chosen as to obtain equally sized strata within each domain, if possible. For the UPS designs with πps, we require that the size variable used to compute the inclusion probabilities is strictly positive. To satisfy this, we used the transformed variable

$$x^*_{dj} = x_{dj} - \min_{\substack{d=1,\ldots,D,\\ j=1,\ldots,N_d}} (x_{dj}) + 1, \tag{5.3.2}$$

in cases where negative values x_{dj} occurred. The model fitted to the sample data was correctly specified for the BHF estimator, i.e. a random intercept model with x_{dj} as the only covariate. For the GREG estimator, we checked whether modelling the domains as random effects improved on not modelling them at all. Since this was not case, which was expected following the arguments in the previous section, we used a model without domain-specific effects. Besides, all the sampling designs considered here are non-informative as x_{dj} is included as a covariate in the model.

The relative biases for the domain estimates are depicted in Figure 5.1. The graph is organised as follows: each of the three green rows corresponds to a fixed finite population generated with the parameter ω. The columns refer to the three estimation methods used. Now, a panel is defined as a particular combination of an estimation method and a certain fixed finite population. Within each panel, four boxplots are displayed, which refer to the performance of the estimator under the different designs used in this study. Thus, each panel allows to compare the impact of the sampling designs on a particular estimator for a certain population. Looking at Figure 5.1, it is obvious that the design-based and model-assisted estimation methods are unbiased in all settings. This is as expected due to the construction of these estimators, which are design-unbiased in the case of the direct estimator and asymptotically unbiased for the GREG estimator. Looking at the right column, we see that depending on the population the unit level EBLUP may suffer from biases in some domains. This result is due to the fact that we condition on a particular realisation of the random intercept model, where the random effect is different from zero.

However, we do not observe a systematic bias in the sense that all estimates are biased in some direction for a particular design or population. Further, within each panel the boxplots are very similar highlighting the non-informativeness of the designs.

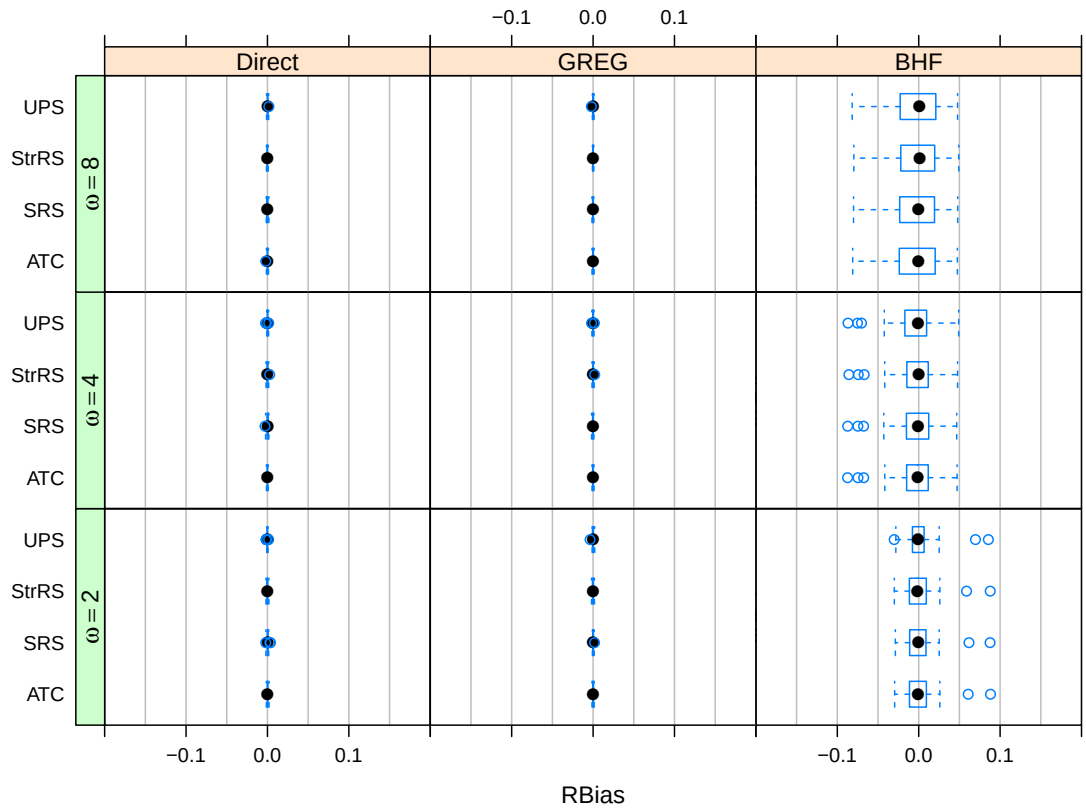

Figure 5.1: Relative biases for planned domains in a quasi-design-based simulation study

To analyse the impact of the sample sizes on the results, we look at the mean absolute relative biases (MARBs) of the domain estimates averaged over domains with the same size. The results are tabulated in Table 5.3. Both the direct and GREG estimators are conditionally unbiased for all sample sizes, populations and sampling designs. In case of the BHF estimator, we see that the conditional absolute bias decreases with the sample size.

An overview of the unconditional RRMSEs is shown in Figure 5.2. It can be seen that for the direct estimator, the SRS design performs in general the worst. Furthermore, we note that in terms of the minimum RRMSE among the domains StrRS performs slightly better for the direct estimator compared to UPS. This changes, however, if we consider the 75% quantile or the maximum RRMSE. The best results with respect to the RRMSE for the direct estimator are clearly obtained when using the ATC scheme. The results for the GREG and the BHF estimators do not vary much with the sampling scheme for a given population. This can be explained by the fact that the assumed model holds and all designs are non-informative. Hence, we would not expect strong design impacts for these procedures. It may be further noted that under the ATC design, the gains from using the GREG instead of the direct estimator are very small. This is due to the symmetric nature of the size variable x_{dj} which is also used as the covariate. Due to its symmetry, the

Table 5.3: MARBs for planned domains in a quasi-design-based simulation study

		Direct			GREG			BHF		
Design	n_d	$\omega = 2$	$\omega = 4$	$\omega = 8$	$\omega = 2$	$\omega = 4$	$\omega = 8$	$\omega = 2$	$\omega = 4$	$\omega = 8$
ATC	6	0.001	0.001	0.001	0.001	0.001	0.001	0.024	0.040	0.035
	14	0.000	0.000	0.001	0.000	0.000	0.000	0.011	0.018	0.032
	30	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.012	0.013
SRS	6	0.001	0.001	0.001	0.001	0.001	0.000	0.024	0.040	0.034
	14	0.001	0.001	0.001	0.000	0.001	0.000	0.011	0.017	0.031
	30	0.001	0.001	0.001	0.000	0.000	0.000	0.009	0.013	0.013
StrRS	6	0.001	0.001	0.000	0.001	0.001	0.000	0.024	0.040	0.034
	14	0.001	0.001	0.001	0.001	0.001	0.000	0.011	0.017	0.031
	30	0.000	0.001	0.000	0.000	0.001	0.000	0.009	0.012	0.013
UPS	6	0.001	0.001	0.001	0.001	0.001	0.001	0.022	0.041	0.034
	14	0.000	0.001	0.000	0.000	0.001	0.000	0.010	0.017	0.030
	30	0.000	0.000	0.000	0.001	0.000	0.000	0.009	0.013	0.013

sample mean $\overline{x}_d$ will be close to the population mean $\overline{X}_d$. Owing to the linear assisting model the GREG estimator can also be expressed as

$$\widehat{\mu}_d^{GREG} = \widehat{\mu}_d^{Direct} + (\overline{X}_d - \overline{x}_d)\widehat{\beta}_1.$$

As $\overline{x}_d \approx \overline{X}_d$ the adjustment term is close to zero and hence the GREG estimator will be almost identical to the direct estimator.

The average relative root mean squared errors (ARRMSEs) conditional on the sample size are given in Table 5.4. It can be seen that for any given population and sample size, ATC yields the lowest ARRMSE for the direct estimator. We further see that StrRS suffers in the very small domains, where it is clearly outperformed by UPS. For a sample size of 14, the difference is not that big anymore and for sample sizes of 30 it performs better than UPS. Furthermore, for ATC designs the direct estimator performs almost as good as the GREG estimator as long as the sample size is not too small. We also see that as the sample size increases, the difference in the RRMSE between the GREG and the BHF estimator decreases for a given population.

Besides the quality of point estimates, we are also interested in the quality of the precision estimates. To do so, we study the average confidence interval coverage rates of 95% confidence intervals averaged at domain sizes. These are presented in Table 5.5. It can be seen that the coverage rates are generally close to the nominal rate for the direct and GREG estimators. For SRS and UPS designs, both these estimation methods are very close to 95 % coverage for all populations, i.e. varying ω and all sample sizes n_d. Regarding the ATC design, the coverage of these design-based and model-assisted procedures is also very good, with some overcoverage when the sample size is very small ($n_d = 6$). With respect to the StrRS design, we note slight undercoverage, when $n_d = 6$ or $n_d = 14$. This may be caused by the fact that the number of strata is quite small for those domains. Looking at the coverage of the BHF estimator we observe some problems, especially in the

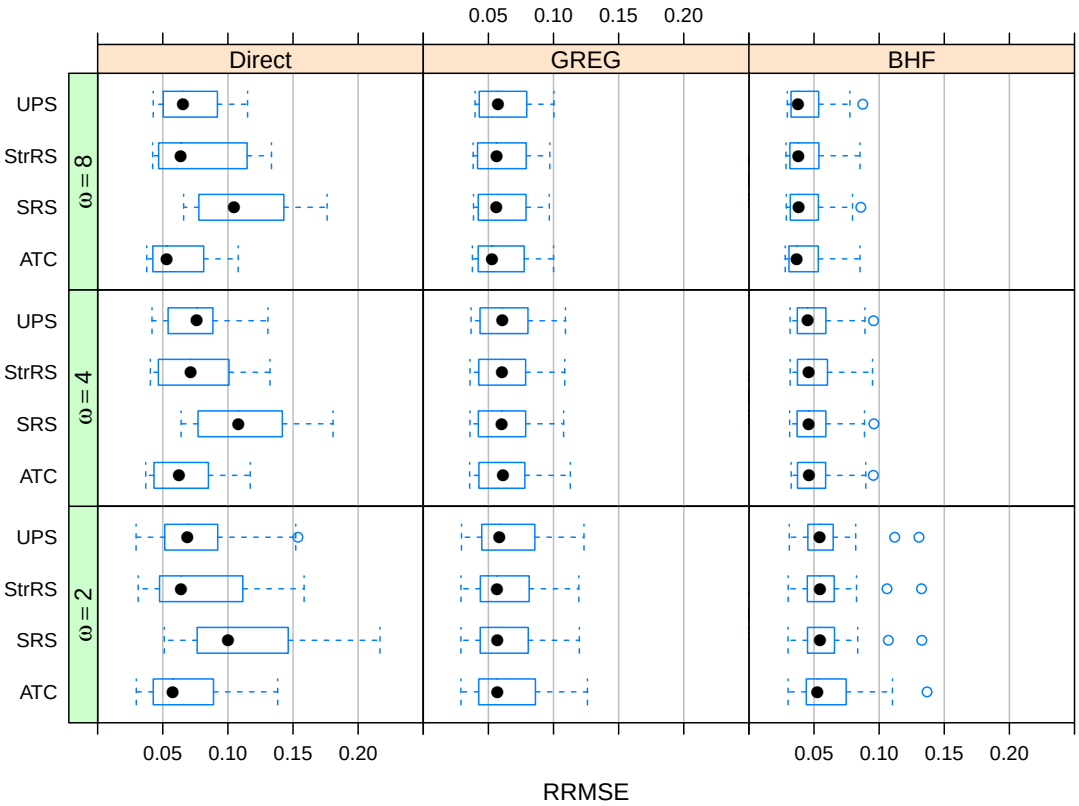

Figure 5.2: RRMSEs for planned domains in a quasi-design-based simulation study

population with $\omega = 8$. It should be taken into account that we used an unconditional MSE estimator based on the decomposition due to Prasad and Rao (1990) to produce confidence intervals, which may be problematic in a quasi-design-based study, where we condition on one realisation of the model. Interestingly, for $\omega = 2$ the coverage rates are very close to the nominal level. In this case, the BHF estimator attaches most weight to the survey regression estimator (see Table 5.2).

A key motivation for the ATC design was to achieve variance reduction for national estimates. Hence, we are also interested in the performance of direct and GREG estimates for national figures under the four sampling designs. A comparison is given in Table 5.6. The columns RBias, RRMSE and ACR denote the relative bias, the relative root mean squared error and the confidence interval coverage of 95% confidence intervals, respectively. In terms of the RRMSE, we note that when a direct estimator is used, the ATC design yields clearly the best estimates in all populations. The RRMSEs under StrRS are in all populations about 10 per cent larger than the ones obtained by ATC, and UPS performs a little worse than StrRS in all populations. The SRS design yields the worst results when applied for the direct estimator with an increase in the RRMSE of about 80 per cent relative to ATC. Note that the differences between designs are generally smaller for the GREG estimator, which can realise variance reductions due to the power of the assisting model. We may further note that the gains of this model may be rather small, if the auxiliary information has been used at the design stage already. The coverage rates of both estimation methods are close to the nominal rate for all populations and all designs.

Table 5.4: ARRMSEs for planned domains in a quasi-design-based simulation study

		Direct			GREG			BHF		
Design	n_d	$\omega = 2$	$\omega = 4$	$\omega = 8$	$\omega = 2$	$\omega = 4$	$\omega = 8$	$\omega = 2$	$\omega = 4$	$\omega = 8$
ATC	6	0.102	0.088	0.087	0.096	0.084	0.082	0.084	0.066	0.049
	14	0.054	0.061	0.054	0.053	0.061	0.053	0.050	0.050	0.047
	30	0.042	0.042	0.042	0.042	0.042	0.042	0.042	0.040	0.035
SRS	6	0.165	0.152	0.154	0.092	0.083	0.084	0.081	0.066	0.050
	14	0.096	0.108	0.104	0.053	0.059	0.055	0.050	0.049	0.048
	30	0.076	0.075	0.074	0.042	0.041	0.041	0.042	0.039	0.034
StrRS	6	0.124	0.110	0.118	0.092	0.083	0.084	0.081	0.065	0.050
	14	0.062	0.070	0.065	0.053	0.060	0.056	0.050	0.049	0.048
	30	0.046	0.046	0.045	0.042	0.041	0.041	0.042	0.039	0.034
UPS	6	0.109	0.099	0.097	0.095	0.084	0.087	0.081	0.066	0.050
	14	0.059	0.074	0.065	0.054	0.060	0.057	0.050	0.048	0.048
	30	0.050	0.052	0.048	0.043	0.042	0.042	0.042	0.040	0.035

5.3.2 Mean estimates under unplanned domains

The example from the previous section highlights the benefits of including the domains at the planning stage, as even design-based or model-assisted estimators may provide reliable domain estimates in this case. In real-life applications, however, small domain estimates are often produced using model-based procedures, whereas a design-based / model-assisted strategy is used to produce aggregate estimates (cf. Little, 2012). This strategy is reasonable, if the domains are unplanned, such that the sample size for the domains is a random variable and may even take the value of zero.

We explore the consequences of unplanned domain structures by means of a quasi-design based simulation study. We use the same populations as in the former section to facilitate a comparison of the results. However, the sampling designs are applied to the population as a whole, ignoring any information about the domain structure. With a total sample size of $n = 500$ this may even yield non-sampled domains in some samples. This causes problems for design-based estimation methods, since a direct estimator is not defined for $n_d = 0$. A question that arises is how to deal with this issue when reporting simulation results. One option would be to exclude the samples in which $n_d = 0$ from the computation of the summary statistics. This approach will not blur the results much, if the number of samples with $n_d = 0$ is small for a particular domain. Another option is to set the point estimates to zero in these cases. This is more in line with the notion of the expected value under a sampling design, which weights all sample estimates with the probability of observing that particular sample. In fact, the sample size for unplanned domains can be modelled by a hypergeometric distribution, which has a non-zero probability for the event that $n_d = 0$. On the other hand this procedure is not very satisfactory in real-life applications. To understand this, imagine that it is known that the variable of interest takes only positive values. In this case, inserting $\widehat{\mu}_d = 0$ does not make sense at all.

In the following, we stick with the second option and insert $\widehat{\mu}_d = 0$ for samples with $n_d = 0$ for both the direct and the MGREG estimator using a random intercept model.

Table 5.5: ACRs for planned domains in a quasi-design-based simulation study

		Direct			GREG			BHF		
Design	n_d	$\omega = 2$	$\omega = 4$	$\omega = 8$	$\omega = 2$	$\omega = 4$	$\omega = 8$	$\omega = 2$	$\omega = 4$	$\omega = 8$
ATC	6	0.959	0.951	0.953	0.961	0.950	0.953	0.953	0.912	0.931
	14	0.950	0.949	0.949	0.950	0.950	0.949	0.959	0.966	0.911
	30	0.949	0.949	0.950	0.950	0.949	0.949	0.952	0.948	0.958
SRS	6	0.949	0.949	0.951	0.950	0.950	0.948	0.961	0.916	0.929
	14	0.950	0.951	0.950	0.951	0.951	0.950	0.957	0.968	0.909
	30	0.950	0.950	0.950	0.949	0.951	0.950	0.951	0.950	0.959
StrRS	6	0.941	0.937	0.938	0.938	0.939	0.938	0.961	0.917	0.932
	14	0.942	0.943	0.940	0.942	0.942	0.940	0.957	0.968	0.912
	30	0.950	0.949	0.949	0.950	0.948	0.949	0.953	0.949	0.959
UPS	6	0.947	0.949	0.946	0.952	0.953	0.951	0.962	0.915	0.933
	14	0.951	0.952	0.950	0.953	0.952	0.950	0.957	0.969	0.918
	30	0.950	0.951	0.951	0.953	0.952	0.950	0.950	0.948	0.959

The MARBs conditional on the expected sample size are tabulated in Table 5.7. While the direct estimator yields very small values for the MARB in all settings, we observe MARB-values around 1.5 to 2 per cent for the MGREG estimator with expected sample sizes below 10. This finding is not very surprising, as the MGREG estimator is asymptotically design-unbiased, which is a large sample property. However, as soon as the sample size increases, the MGREG estimator displays very good results in terms of the MARB. For the model-based BHF estimator, we see relatively high values of the MARB, especially with small sample sizes. This is a result of taking one realisation of the model as the fixed and finite population.

The ARRMSEs conditional on the expected sample size are tabulated in Table 5.8. Compared to the planned domain case illustrated in Table 5.4, the much higher ARRMSEs of the direct estimator are striking. Furthermore, for this estimator, now the UPS design yields the best results, with all other designs leading to similar results. Moreover, for the MGREG estimator, we observe a deterioration of the quality of the estimates versus the planned domain case for the smallest domains with the GREG. However, for the largest domains with an expected sample size of 30 and more slightly better results can be obtained, owing to the increase of the expected sample size versus the planned domain case. Finally, the BHF estimator does not suffer from unplanned domains, which was also expected.

The quality of estimates for the national mean is given in Table 5.9. The relative biases of the estimators under all scenarios is nearly zero and is hence not reported in this table. We note that ATC and StrRS designs yield the best results for the direct estimator in terms of the RRMSE. The results under UPS are worse due to the presence of the non-zero intercept term. For the GREG estimator under a fixed-effects assisting model we hardly observe any impact of design, which can be attributed due to the correctly specified model which does not leave substantial room for improvement over SRS. The same can be said about the MGREG, which performs better in general due to the additional random intercept and the unplanned domain structure. In terms of the confidence interval coverage

Table 5.6: National estimates with planned domains in a quasi-design-based simulation study

		Direct			GREG		
Design	ω	RBias	RRMSE	ACR	RBias	RRMSE	ACR
ATC	2	0.0000	0.0104	0.9464	0.0000	0.0102	0.9467
	4	0.0000	0.0107	0.9486	0.0000	0.0105	0.9482
	8	0.0000	0.0106	0.9458	0.0000	0.0105	0.9468
SRS	2	0.0001	0.0185	0.9481	0.0000	0.0103	0.9460
	4	0.0002	0.0190	0.9496	0.0000	0.0105	0.9480
	8	0.0001	0.0188	0.9502	0.0001	0.0106	0.9481
StrRS	2	−0.0001	0.0113	0.9549	−0.0001	0.0100	0.9544
	4	0.0001	0.0120	0.9488	0.0001	0.0106	0.9460
	8	−0.0001	0.0117	0.9510	0.0000	0.0104	0.9481
UPS	2	0.0000	0.0120	0.9515	0.0000	0.0105	0.9493
	4	0.0001	0.0129	0.9577	0.0001	0.0105	0.9531
	8	0.0000	0.0121	0.9542	0.0000	0.0107	0.9505

rates, we observe good results for the direct estimator. The constant undercoverage of the MGREG is due to the fact that a simple residual variance estimator has been used. Whilst this procedure can be justified under a planned domain structure, it ignores some of the complications arising with unplanned domains as outlined in Section 2.2. Note that for a GREG estimator based on a linear fixed effects model, the residual variance estimator works well. Thus, with unplanned domains including domain-specific random effects in the assisting model makes a huge difference.

5.3.3 Mean estimates under a misspecified model

In this section, we analyse the impact of model misspecification on the quality of small area estimates for the designs discussed in the previous section. The misspecification considered takes the form of an omitted variable bias, i.e. a relevant covariate is not included in the analysis model. In detail, the finite populations were constructed as realisation of the random intercept model with two auxiliary variables, parametrised as

$$\begin{aligned} y_{dj} &= 6 + 3 \cdot x_{dj1} + 2 \cdot x_{dj2} + v_d + \varepsilon_{dj}, \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D, \\ v_d &\overset{i.i.d}{\sim} N(0, 2) \\ \varepsilon_{dj} &\overset{i.i.d}{\sim} N(0, 4). \end{aligned} \tag{5.3.3}$$

The covariates x_{dj1} and x_{dj2} were independently drawn as $x_{dji} \overset{i.i.d.}{\sim} N(1, 1)$, $i = 1, 2$. Furthermore, x_{dj1} was used as the design variable in ATC, StrRS and UPS designs. For UPS designs, we used the same linear transformation (5.3.2) as in the previous section to ensure non-negativity. The sampling design treats the domains as planned, i.e. on the first stage the domains are strata. On the second stage, ATC, SRS, StrRS and UPS designs are employed within domains.

Table 5.7: MARBs for unplanned domains in a quasi-design-based simulation study

		Direct			MGREG			BHF		
Design	E(n_d)	$\omega = 2$	$\omega = 4$	$\omega = 8$	$\omega = 2$	$\omega = 4$	$\omega = 8$	$\omega = 2$	$\omega = 4$	$\omega = 8$
ATC	0 - 10	0.004	0.003	0.006	0.018	0.020	0.018	0.051	0.040	0.031
	10 - 30	0.002	0.002	0.003	0.001	0.002	0.001	0.013	0.025	0.025
	30 +	0.001	0.001	0.002	0.000	0.000	0.000	0.004	0.003	0.013
SRS	0 - 10	0.005	0.003	0.004	0.018	0.020	0.018	0.051	0.040	0.031
	10 - 30	0.002	0.001	0.002	0.000	0.001	0.001	0.012	0.025	0.025
	30 +	0.002	0.002	0.001	0.000	0.000	0.000	0.004	0.002	0.013
StrRS	0 - 10	0.003	0.005	0.003	0.018	0.020	0.017	0.051	0.040	0.031
	10 - 30	0.002	0.003	0.002	0.001	0.002	0.001	0.013	0.025	0.025
	30 +	0.001	0.002	0.002	0.000	0.000	0.000	0.004	0.003	0.013
UPS	0 - 10	0.004	0.005	0.003	0.016	0.020	0.018	0.049	0.040	0.031
	10 - 30	0.002	0.001	0.002	0.001	0.002	0.001	0.013	0.025	0.025
	30 +	0.001	0.001	0.001	0.000	0.000	0.000	0.004	0.003	0.013

In a similar vein to Lehtonen and Veijanen (2009, Section 5.2), we considered three kinds of models which were fitted to the sample data. In model A, only x_{dj1} was included as a covariate, while in model B only x_{dj2} was included in the model fitting and for model C both x_{dj1} and x_{dj2} were used as auxiliary information. Hence, models A and B are subject to model misspecification as a relevant predictor is not included in the specification. Furthermore, for model B the sampling design is potentially informative, as the variable determining the sample selection is not included in the model. The assisting model for the GREG is based on a linear fixed effects model, where in the case of an UPS design we included both the design weights and the domain membership as an additional covariate. This has been done in order to deal with the potential informativeness of a type B model and due to the fact that with an UPS scheme $\widehat{N}_d = N_d$ is generally not satisfied for a particular sample. The different models are summarised for convenience in Table 5.10.

As a simulation setting, we consider a quasi-design-based setup. We take one realisation of (5.3.3) as the fixed and finite population from which $R = 10{,}000$ samples are drawn. A similar strategy was employed by Lehtonen et al. (2003, section 6.2). Additionally, we conducted a finite population model-based simulation study which yielded similar conclusions. Those results are reported in the Appendix C.

The results in terms of the mean absolute relative biases for the domain estimates in the quasi-design-based setup are given in Table 5.11. Note that for the direct estimator, we do not report a model, as this procedure does not make use of a statistical model. It can be seen that both the direct and GREG estimator do not exhibit biases irrespective of the specification of the model, the survey design or sample size. With respect to the model-based BHF estimator we observe the highest absolute biases in model B, where we encounter both model misspecification and for the UPS design, an informative selection. Under this model, SRS yields the lowest biases for all groups of domains. While the results under ATC and StrRS designs are very close, the average absolute bias increases for UPS. For non-informative models A and C we do not observe a strong impact of the sampling design on the biases. Thus, applying a misspecified model without including the

Table 5.8: ARRMSEs for unplanned domains in a quasi-design-based simulation study

		Direct			MGREG			BHF		
Design	E(n_d)	$\omega = 2$	$\omega = 4$	$\omega = 8$	$\omega = 2$	$\omega = 4$	$\omega = 8$	$\omega = 2$	$\omega = 4$	$\omega = 8$
	0 - 10	0.463	0.465	0.465	0.144	0.141	0.141	0.092	0.075	0.051
ATC	10 - 30	0.259	0.260	0.257	0.054	0.057	0.054	0.050	0.055	0.046
	30 +	0.161	0.163	0.162	0.032	0.036	0.034	0.031	0.033	0.032
	0 - 10	0.465	0.463	0.465	0.145	0.141	0.140	0.092	0.075	0.051
SRS	10 - 30	0.259	0.261	0.258	0.054	0.057	0.054	0.049	0.055	0.046
	30 +	0.161	0.163	0.163	0.032	0.036	0.034	0.031	0.033	0.032
	0 - 10	0.465	0.462	0.467	0.145	0.141	0.141	0.092	0.075	0.051
StrRS	10 - 30	0.259	0.260	0.258	0.054	0.057	0.054	0.050	0.055	0.046
	30 +	0.161	0.163	0.162	0.032	0.036	0.034	0.031	0.033	0.032
	0 - 10	0.442	0.444	0.448	0.143	0.144	0.143	0.091	0.075	0.051
UPS	10 - 30	0.246	0.249	0.249	0.055	0.058	0.055	0.050	0.055	0.046
	30 +	0.155	0.156	0.156	0.033	0.037	0.035	0.031	0.033	0.032

size variable (model B) to data gathered by an UPS scheme is not a good idea.

The ARRMSEs for the different groups of domain sizes in the quasi-design-based study are given in Table 5.12. It can be seen that for the direct estimator, SRS yields the largest ARRMSE in all domain size groups. Furthermore, ATC and UPS yield slightly better estimates than StrRS. Looking at the GREG estimator, we see that the sampling scheme is not an issue for A and C models. For model B, however, the impact is very strong. In the baseline setting with SRS applied within domains, we achieve a ten to 15 per cent reduction compared to the direct estimates for all size groups. Roughly similar gains are obtained under the UPS scheme, where a more complex assisting model was entertained. Under StrRS, the efficiency gains achieved over the direct estimator are in the range of 20 to 30 per cent and it is also a far better sample scheme under this model than SRS or StrRS. Clearly the best choice under model B for the GREG estimator is the proposed ATC scheme. It is striking that applying the B-model with ATC yields virtually the same results as in the case of the correctly specified C-model. This can be explained by the fact that the clusters are constructed based on the omitted variable x_{dj1}. With ATC the sample mean of x_{dj1} will be very close to the population mean such that in a C-model, the regression coefficient on x_{dj1} will not contribute much to the regression correction term. Focussing on the BHF estimator, we do not observe an impact of the sampling scheme in models A and C. For model B, the squared bias under UPS leads to a deterioration of the ARRMSE. Moreover, both StrRS and ATC schemes lead to a reduction in the ARRMSE compared to the SRS scheme in model B. In particular, for smaller domains with $n_d = 6$ or $n_d = 14$ ATC performs even better than StrRS.

The ACRs are shown in Table 5.13. The nominal coverage rate of 95 per cent is met for the direct estimator, except for $n_d \in \{6; 14\}$ for the StrRS design. This finding holds for the GREG estimator as well, where we observe a slight undercoverage when the UPS design is used. For the BHF estimator, the confidence intervals are slightly conservative under models A and C, except for very small domains. For the B-model however, the actual coverage rates differ from the nominal ones by far for ATC, StrRS and UPS designs, unless

Table 5.9: National estimates with unplanned domains in a quasi-design-based simulation

		Direct		GREG		MGREG	
Design	ω	RRMSE	ACR	RRMSE	ACR	RRMSE	ACR
ATC	2	0.0105	0.9493	0.0105	0.9490	0.0091	0.9386
	4	0.0110	0.9484	0.0110	0.9487	0.0104	0.9346
	8	0.0102	0.9455	0.0102	0.9454	0.0099	0.9358
SRS	2	0.0173	0.9529	0.0106	0.9473	0.0092	0.9376
	4	0.0186	0.9513	0.0107	0.9510	0.0102	0.9384
	8	0.0175	0.9513	0.0102	0.9475	0.0099	0.9354
StrRS	2	0.0106	0.9506	0.0105	0.9504	0.0091	0.9375
	4	0.0110	0.9497	0.0108	0.9493	0.0101	0.9401
	8	0.0103	0.9486	0.0101	0.9499	0.0098	0.9398
UPS	2	0.0118	0.9503	0.0109	0.9487	0.0096	0.9348
	4	0.0127	0.9504	0.0110	0.9492	0.0104	0.9384
	8	0.0120	0.9492	0.0102	0.9478	0.0099	0.9388

Table 5.10: Models for the simulation study with misspecification

	Covariates	Informativeness (UPS)
A	x_{dj1}	no
B	x_{dj2}	yes
C	x_{dj1} and x_{dj2}	no

$n_d = 14$. As the coverage rates are far below the nominal rate for $n_d = 6$, and for the ATC and SRS designs far above the nominal rate for $n_d = 30$, the behaviour is not convincing. In the case of the SRS design, however, the coverage is very good, except for the case with $n_d = 6$.

The quality of the national estimates in the design-based simulation study are given in Table 5.14. Looking at the direct estimates, we note that in terms of the national RRMSE the ATC design yields the best estimates. Furthermore, the StrRS and UPS designs give very similar results. The SRS scheme yields unambiguously the worst results. For the GREG estimator we do not observe major differences when the assisting model is correctly specified. If the design variable is included in the model, the results are not much different either, but the UPS scheme yields slightly worse results. In case the B-model is used, the sampling designs makes a huge difference. Applying StrRS yields a RRMSE which is only 10 percentage points higher than if the correct model is used. If ATC is used, the results can hardly be distinguished from the correct specification. For UPS and StrRS designs, however, much higher RRMSEs result.

5.3.4 Estimating poverty rates

Another application for which ATC could be suitable is the estimation of statistics which depend as well on the median rather than the mean alone. Examples in this regard are poverty measures such as the at-risk-of-poverty rate (ARPR) ot the relative poverty gap.

Table 5.11: MARBs under model misspecification for planned domains in a quasi-design-based simulation study

		Direct	GREG			BHF		
Design	n_d		A	B	C	A	B	C
ATC	6	0.001	0.001	0.001	0.001	0.052	0.092	0.035
	14	0.001	0.000	0.000	0.001	0.025	0.047	0.013
	30	0.000	0.000	0.000	0.000	0.013	0.024	0.006
SRS	6	0.001	0.001	0.001	0.001	0.052	0.075	0.035
	14	0.001	0.000	0.001	0.000	0.025	0.035	0.013
	30	0.001	0.000	0.001	0.001	0.013	0.017	0.006
StrRS	6	0.001	0.001	0.001	0.000	0.052	0.092	0.035
	14	0.001	0.001	0.001	0.000	0.025	0.049	0.013
	30	0.000	0.000	0.000	0.000	0.013	0.022	0.006
UPS	6	0.001	0.001	0.002	0.001	0.052	0.112	0.035
	14	0.001	0.001	0.002	0.000	0.025	0.069	0.013
	30	0.000	0.000	0.001	0.000	0.013	0.043	0.007

These statistics are based on a comparison of a welfare variable, e.g. income, with the poverty line, which is typically estimated as 60 per cent of the median of the equivalised household income.

In the following we compare ATC with SRS, both of which are conducted within areas, for the purpose of estimating the ARPR. Among the estimators we consider design-based methods as well as the model-based EBP due to Molina and Rao (2010) introduced in Section 3.3. While the potential of ATC has been introduced for design-based methods already, its benefits for model-based predictors may be less obvious. Indeed, ATC cannot lead to a better modelling if the underlying model is correctly specified. It may, however, yield more precise estimates of the poverty line, especially when the threshold is estimated for small domains rather than the country as a whole.

We present results from a quasi-design-based study where the universe consists of $N = 30000$ units belonging to $D = 30$ domains. In the same fashion as in Section 4.4, the units were allocated to domains following Lehtonen and Veijanen (2009, Section 5.2) with probabilities proportional to $\exp(q_d)$ where $q_d \overset{ind}{\sim} U[0, 2.9]$. We consider three types of sample sizes presented in Table 5.15. As can be seen from Table 5.15, the sample sizes are in general larger than those used in other simulation studies in preceding chapters. This is due to the fact that if local poverty thresholds are to be used, we would not want to estimate them based on only 6 or 14 observations. Moreover the variation of the sample sizes permits us to study the influence of different sample sizes on the simulation results. The population is drawn once as a realisation from the following model:

$$\begin{aligned} \log(y_{dj}) &= 1 + 1 \cdot x_{dj} + v_d + e_{dj}, \quad j = 1, \ldots, N_d \quad d = 1, \ldots, D, \\ v_d &\overset{i.i.d}{\sim} N(0, 0.3^2) \\ e_{dj} &\overset{i.i.d}{\sim} N(0, 0.6^2). \end{aligned} \tag{5.3.4}$$

Table 5.12: ARRMSEs under model misspecification for planned domains in a quasi-design-based simulation study

		Direct	GREG			BHF		
Design	n_d		A	B	C	A	B	C
ATC	6	0.114	0.108	0.075	0.075	0.087	0.100	0.069
	14	0.065	0.064	0.045	0.045	0.057	0.057	0.042
	30	0.042	0.042	0.030	0.030	0.039	0.035	0.029
SRS	6	0.148	0.105	0.130	0.074	0.085	0.106	0.069
	14	0.096	0.066	0.084	0.046	0.058	0.070	0.043
	30	0.062	0.042	0.054	0.030	0.039	0.048	0.029
StrRS	6	0.121	0.105	0.097	0.075	0.085	0.104	0.069
	14	0.072	0.066	0.053	0.047	0.057	0.062	0.043
	30	0.045	0.042	0.033	0.030	0.039	0.035	0.029
UPS	6	0.111	0.106	0.089	0.076	0.085	0.131	0.069
	14	0.070	0.067	0.062	0.047	0.057	0.091	0.043
	30	0.044	0.043	0.040	0.031	0.039	0.061	0.029

In (5.3.4), the logarithm of the dependent welfare variable y_{dj} is modelled via a nested error regression model and independence between the v_d and e_{dj} is assumed. The covariate, which is also used to draw the ATC sample is taken as independently and identically distributed realisations of a $\mathcal{U}[0,1]$ random variable.

The relative biases of the predictor due to Molina and Rao (2010) and the direct estimator conditional on the domain-specific sample sizes are given in Figure 5.3. The rows of this graph correspond to the different domain-specific sample sizes and are ordered in an increasing fashion from the bottom to the top. We observe relatively large boxes for the direct estimator in domains with $n_d = 50$, which are much more narrow in the cases of $n_d = 100$ or $n_d = 150$. Moreover, striking differences between the SRS or ATC design can not be found. For the MR estimator, a tendency to overestimate the true ARPR can be seen for small domains, with little further distinction between the two sampling designs.

The RRMSEs conditional on the sample sizes are presented in Figure 5.4. Especially for the direct estimator, we see a strong impact of the sample size on the RRMSEs. The impact of the sample size is also clearly visible for the MR estimator. The performance of the direct estimator is very similar in both sampling designs. Interestingly, a strong impact of the design on the performance of the model-based MR estimator can be seen. In this case, we observe within each panel smaller RRMSEs when the ATC method is applied compared to SRS.

5.4 Summary and discussion

In this chapter, we introduced the ATC method, which allows to realise variance reductions for a design-based estimation method versus SRS, while leading to a self-weighting scheme on the level at which it is applied. Therefore, the proposed method does not lead

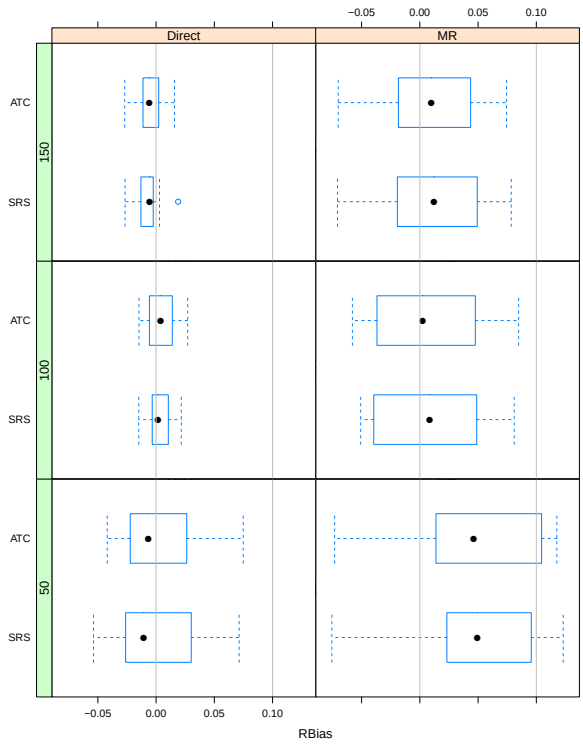

Figure 5.3: Relative biases for the ARPR in a quasi-design-based simulation study

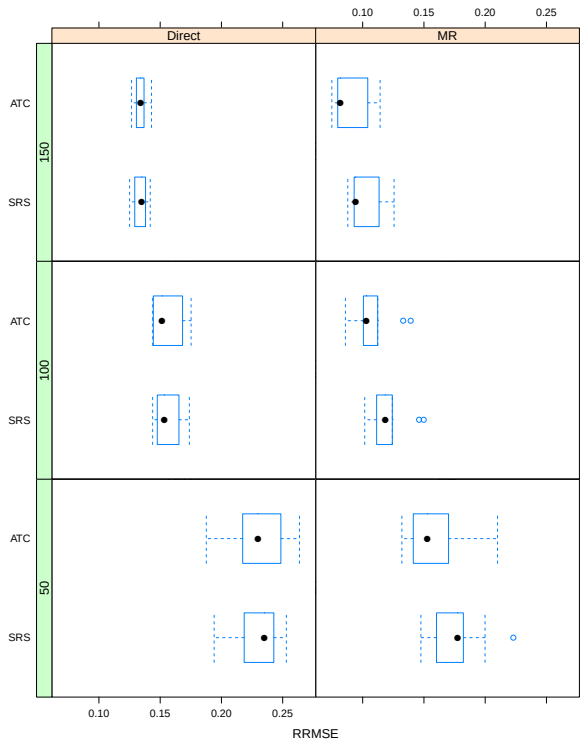

Figure 5.4: RRMSEs for the ARPR in a quasi-design-based simulation study

Table 5.13: ACRs under model misspecification for planned domains in a quasi-design-based simulation study

		Direct	GREG			BHF		
Design	n_d		A	B	C	A	B	C
ATC	6	0.950	0.948	0.954	0.957	0.914	0.794	0.929
	14	0.952	0.952	0.950	0.949	0.958	0.956	0.963
	30	0.949	0.949	0.950	0.951	0.956	0.993	0.956
SRS	6	0.948	0.952	0.949	0.953	0.920	0.878	0.930
	14	0.950	0.950	0.950	0.950	0.955	0.949	0.959
	30	0.949	0.950	0.950	0.949	0.957	0.958	0.954
StrRS	6	0.938	0.941	0.940	0.940	0.919	0.807	0.930
	14	0.941	0.941	0.941	0.941	0.957	0.948	0.959
	30	0.948	0.949	0.948	0.948	0.956	0.991	0.955
UPS	6	0.951	0.947	0.948	0.946	0.917	0.804	0.929
	14	0.951	0.947	0.949	0.945	0.955	0.860	0.959
	30	0.951	0.947	0.938	0.946	0.956	0.890	0.953

to impediments for model-based estimates which may be needed on lower levels of aggregation. Besides, we showed that ATC is the optimal choice for minimising the variation between clusters amongst all schemes, which construct clusters of size 2 and where one unit with a value below the median of the clustering variable is coupled with a unit with a value above this median. Assuming a linear model, a mild condition under which the proposed strategy in expectation yields better results than a sample collected by SRS was established. Furthermore, we detailed how the same condition arises under a simple small area model (5.2.27) when the ATC scheme is applied within domains. The beauty of this result is that the superiority of the ATC scheme for domains automatically translates to the national mean. We also developed approximations which are applicable when a small area model governs the relationship, but the domains are unplanned.

We conducted simulation studies under the model (5.2.27) to study the impact of planned versus unplanned domain structures. The study indicates that while the quality of domain estimates using a model-based EBLUP is hardly affected by unplanned domain structures, a severe deterioration can be seen for direct and indirect, model-assisted estimation methods. A further study analysed the impact of a potential model misspecification on the results for domain and national estimates. The findings suggest that the EBLUP can compensate to some extent for the omission of a relevant predictor as long as the sampling mechanism is not informative. Furthermore, while the asymptotic unbiasedness of the GREG estimator is unaffected by the model choice, its efficiency clearly depends on choosing the right model. When a misspecified model is used and the variable which determines the design is not among the covariates, the sampling selection method matters and our results indicate potential advantages of the ATC approach. Moreover, we also addressed the estimation of the ARPR when locally estimated thresholds are used. The simulation study indicated that the predictor proposed by Molina and Rao (2010) benefits from a more precise estimation of the poverty line due to the ATC scheme.

Obviously, a requirement for the ATC method to be non-informative is that the sampling

Table 5.14: National estimates under misspecification in a quasi-design-based simulation study

		Direct			GREG		
Design	Model	RBias	RRMSE	ACR	RBias	RRMSE	ACR
ATC	A	0.0000	0.0111	0.9524	0.0001	0.0111	0.9482
	B				0.0000	0.0078	0.9545
	C				0.0000	0.0079	0.9469
SRS	A	−0.0002	0.0162	0.9505	0.0000	0.0111	0.9500
	B				−0.0003	0.0140	0.9493
	C				0.0000	0.0079	0.9492
StrRS	A	−0.0001	0.0119	0.9488	0.0000	0.0110	0.9503
	B				0.0000	0.0089	0.9534
	C				0.0000	0.0079	0.9512
UPS	A	0.0001	0.0115	0.9529	−0.0001	0.0115	0.9459
	B				−0.0010	0.0114	0.9341
	C				0.0000	0.0081	0.9471

Table 5.15: Sample sizes for the quasi-design-based simulation study of the ARPR

d	Range of N_d	n_d
$1, \ldots, 10$	$149 - 411$	50
$11, \ldots, 20$	$417 - 1238$	100
$21, \ldots, 30$	$1276 - 2536$	150

design on higher levels is ignorable as well. This assumption may be violated if the two-stage design illustrated in Algorithm 2 is employed, and the sampling of areas is already informative. The ATC method is very simple to implement, provided an auxiliary variable is available from the register which is known to be related to the variable of interest. This requirement is no more demanding than the assumptions needed to employ sampling with probabilities proportional to size. Moreover, the direction of the correlation between the variable of interest and the auxiliary design variable z_k is irrelevant for the proposed design, as long as a linear relationship seems plausible. For πps designs, however, a negative correlation between the variable of interest and the auxiliary variable may lead to a substantial increase of the variance relative to SRS. Hence, ATC provides additional robustness regarding the implicitly assumed model.

There are also other options which may yield variance reductions compared to SRS such as StrRS. Amongst these are approaches towards optimal model-based stratification for the GREG estimator as discussed in Särndal et al. (1992, Section 12.4), which require knowledge about the error structure of the assisting regression model and a rule to determine the stratum membership. Thus, the survey planner needs a comprehensive knowledge about the model, which is by far more demanding than knowing the values of some auxiliary variable. Nonetheless, in some applications, the required level of knowledge using optimal model-based stratification may be less of an issue (cf. Särndal et al., 1992, Section 12.5 for examples). Despite these considerations, the probably most commonly known appli-

cations of StrRS schemes are in situations where the strata are (pre-) defined using the values of one or more categorical variables. While the proposed method is not directly applicable in these circumstances, a simple solution is to use StrRS on the first stage and then within strata ATC at the second stage. One example in this regard is precisely when small area estimates are desired and the areas of interest are known beforehand, such that they can be treated as planned domains in sampling scheme. Both the developments in Section 5.2.3 as well as the simulation results in Section 5.3.1 indicate the usefulness of this approach.

Another issue which remains open for further research, is how to extend the ATC approach to cases with more than auxiliary variable used in determining the antithetic clusters. This is not straightforward, since there is no undisputed multivariate ordering statistic, such as the rank or the quantile in the univariate case. If the other variables used to determine the cluster membership are also continuous, the method of principal components could be a practical choice. A potential solution in this regard is laid out in Algorithm 4.

Algorithm 4 Possible extension of ATC for multiple design variables

1. Apply a principal component analysis to the standardised matrix of design auxiliary information to obtain uncorrelated principal components.
2. Order the matrix of the principal components such that the entries are listed in decreasing order of their Euclidean distance to the origin.
3. Compute the Euclidean distance from the first data point of the permuted principal components matrix to all other data points. Choose the first cluster, such that the entry having the largest distance from the first entry is coupled with the first entry.
4. Remove the already coupled entries from the permuted matrix in step 2.
5. Repeat steps 3 and 4 until all clusters are formed.

Chapter 6

Selected applications

Throughout our work, different strategies towards producing accurate small area statistics have been presented. Among them is the incorporation of domains into the sampling design which together with design-based or model-assisted estimation approaches yields precise estimates for large domains, covered in Chapter 2. These strategies further comprise the use of model-based estimation approaches for small domains that have been presented in Chapters 3 and 4. In this chapter, we test the different strategies in design-based simulations to investigate their performance in situations where the structure of the data is unknown beforehand.

In Section 6.1, we apply domain estimation in a business survey, where the variable of interest is strongly right-skewed, such that an appropriate transformation is needed to find a suitable model. Moreover, various allocation procedures of the sample size to the strata are compared. Furthermore, we use data from the Luxembourgian Census 2011 and the actual sampling design of the Luxembourgian labour force survey to compare several estimation approaches for totals of employed and unemployed in socio-demographic classes in Section 6.2. In Section 6.3, we analyse the impact of different stratification schemes and allocation procedures on various domain estimation approaches for poverty rates.

6.1 Small area estimation for business data

6.1.1 Setup

In this application, we compare several allocation mechanisms for StrRS as well as different small area estimation strategies on a close to reality data set based on Italian business data. The aim of our study, which is based on the TRItalia data set described in Bernardini Papalia et al. (2013, Chapter 5), is to estimate the mean of the labour costs in each domain. TRItalia is based on the Italian register of enterprises, ASIA, where the attention is restricted to the small and medium enterprises with less than 100 employees.[1] In business surveys, the domains of interest are frequently composed as combinations of geographical information with the industry sector. To follow this convention, we determined

[1]We kindly thank ISTAT for providing the ASIA archive as well as the PMI survey of small and medium enterprises.

the domains as cross-classifications of NUTS 1 regions and the first digit of the industry classification, which leads to $D = 45$ domains. The population was further stratified within each domain according to the company size in terms of the number of employees. As most enterprises in the data set have less than 5 employees, we aggregate the size variable into two groups, where the first group comprises enterprises with 1 to 5 employees, whereas all other enterprises belong to the second group. Note that this setup describes a planned domain structure due to the nesting of the strata within domains.

A list of the sampling designs used in this study is given in Table 6.1. In order to facilitate comparisons, we first determined the allocation due to Choudhry et al. (2012), specifying a 10 per cent tolerance on coefficients of variation for the strata means and 1.5 per cent on the coefficient of variation of the national mean. This led to a sample size of $n = 67{,}989$, which was subsequently used for the other allocation procedures as well. Since we know the universe in this study, we could actually use the stratum-specific standard deviations, which are, of course, unknown in practice. This total sample size is also quite close to the actual sample size used in the Italian small and medium enterprises survey from 2003. The stratum sizes varied between 799 and 364,294 enterprises whereas for the domains the range is from 6340 to 398,874 units. Since our analysis is based on the TRItalia data set, this setting amounts to a design-based simulation study, where we chose the number of Monte-Carlo replications to be $R = 10{,}000$.

Table 6.1: Allocation procedures for the simulation study on Italian business data

Abbreviation	Description
COSTA	Costa-type allocation (2.4.1) with $c = 0.5$
CRH	NLP approach based on (2.4.5)
EQ	Equal allocation (2.3.5)
OPT	Optimal allocation (2.3.9)
PROP	Proportional allocation (2.3.4)

The estimators used in this study are summarised in Table 6.2. For the variable of interest, we calculated a skewness coefficient in the population of 34.86, implying that it is heavily right-skewed. Therefore, we decided to focus on models using the log transformation of the labour costs as our dependent variable. The covariates we considered comprise the standardised log of the total number of employees as well the legal status of the enterprise. A more complex model could not be entertained as the number of covariates present in the data set which can be assumed to be known at the design stage is scarce. We also considered using the EBP under the augmented model given by (4.3.5) or the SWEE predictor described in Section 4.3, but residual plots against different choices of the augmenting variable did not indicate informativeness in any design. Hence, we omitted both these predictors from the subsequent analysis. For the ALLOG predictor, we modelled the log of the direct estimates. As discussed in Section 3.4, we approximate the variances of the log of the direct estimates via Taylor-linearisation as $\sigma^2_{\varepsilon,d} = \psi_d / \left(\widehat{\mu}_d^{Direct}\right)^2$. MSE estimates for the ALLOG and the EBP were obtained by means of the jackknife approach due to Jiang et al. (2002). For the benchmarked predictor, we did not consider MSE estimation. The variance estimates for both the GREG and direct estimators were obtained using standard formulae valid for stratified random sampling. To compute the confidence intervals, we used the normal method, which leads to the following intervals:

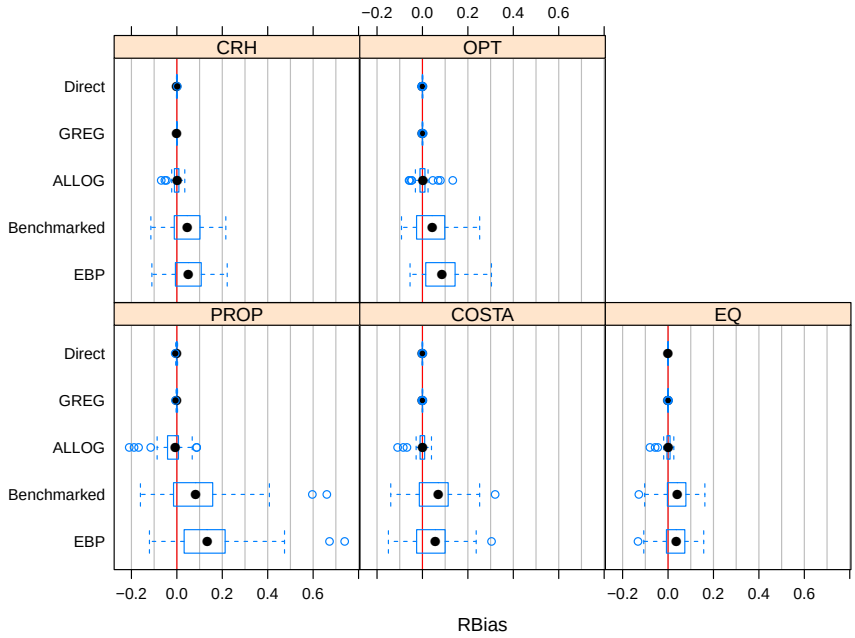

Figure 6.1: Relative biases of small area estimators for the simulation study on Italian business data

$$\left[\widehat{\mu}_d - z_{1-\alpha/2} \cdot \sqrt{\mathrm{MSE}(\widehat{\mu}_d)}; \widehat{\mu}_d + z_{1-\alpha/2} \cdot \sqrt{\mathrm{MSE}(\widehat{\mu}_d)}\right], \tag{6.1.1}$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$-quantile of the standard normal distribution.

Table 6.2: Estimators for the simulation study on Italian business data

Abbreviation	Description
EBP	EBP defined in (3.4.10)
Benchmarked	Benchmarked predictor (4.5.1)
ALLOG	Predictor (3.4.4) under a log-transformed area level model
GREG	Indirect estimator, predictions were obtained from the unit level lognormal mixed model
Direct	A direct estimator of the domain mean

6.1.2 Results for domain estimation

The relative biases are depicted in Figure 6.1. We note that both the direct and the GREG estimator do not suffer from Monte-Carlo biases in any design. Furthermore, the ALLOG predictor also yields reliable results, except for the optimal and the proportional allocation. Worse results occur for the EBP and the benchmarked predictor, which both suffer especially under the proportional allocation.

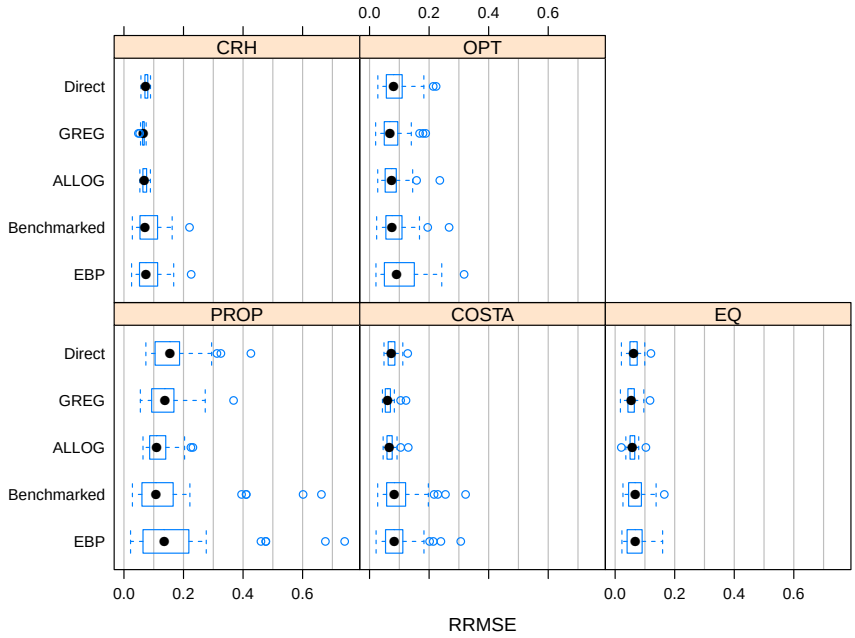

Figure 6.2: RRMSEs of small area estimators for the simulation study on Italian business data

The results in terms of the RRMSEs are shown in Figure 6.2. We note that, in general, the proportional allocation yields the worst results for domain estimation. In the case of the benchmarked predictor and the EBP this is largely due to the contribution of the squared bias. With regards to the direct and GREG estimators, we note some advantages for the latter, which implies that the assisting model has some predictive power. Overall, the most convincing results under proportional allocation are obtained by the model-based estimator under the area level model, which yields by far the lowest maximum RRMSE. Better results in terms of the RRMSE are obtained by the other designs, of which the optimal allocation is the least convincing for domain estimation. The best results for domains emerge under the CRH allocation (2.4.5) and the equal allocation (2.3.5). It is interesting to see that under the equal allocation, none of the estimation methods yields a maximum RRMSE higher than 20 per cent. In this case, we see a slight advantage of the ALLOG estimator versus the GREG estimator. Under the CRH allocation, the unit level model-based procedures both suffer a little, but very good results are obtained using the direct, GREG and ALLOG estimators. Owing to the fact that the upper boundary on the stratum coefficients of variation was taken to be 10 per cent for the direct estimator, the maximum RRMSE is well below 10 per cent for any domain using this procedure. The GREG estimator can improve on these results owing to the predictive power of the assisting model, whereas the model-based ALLOG realises some RRMSE reduction by modelling the direct estimates.

The average confidence interval coverage rates for the estimators except the benchmarked estimator are reported in Table 6.3. It is striking that under no design any estimator meets the nominal rate of 95 per cent. In the case of the EBP, this result is not surprising regarding the systematic biases apparent from Figure 6.1, which are not accounted for in the MSE estimation. For the three other estimators, we observe coverages of 92-93 per cent for all designs except the proportional allocation. In the case of the ALLOG

Table 6.3: Average coverage rates of small area estimators for the simulation study on Italian business data

	Designs				
Estimator	COSTA	CRH	EQ	OPT	PROP
EBP	0.490	0.586	0.555	0.578	0.394
ALLOG	0.921	0.931	0.929	0.930	0.867
GREG	0.925	0.926	0.928	0.925	0.918
HT	0.920	0.923	0.928	0.918	0.857

predictor, this undercoverage can be explained by the fact that the model assumptions, while not completely implausible, are not fully met. The undercoverages of the GREG and direct estimators require further investigation.

In Figure 6.3, we plot the average coverage rates of the GREG and direct estimators against the domain-specific skewness coefficients in the population. A striking aspect is the wide range of skewness coefficients over the domains, which vary from about 10 to over 60. Furthermore, we use different colours for each point, depending on the sample size in each domain. Blue points indicate domains where $n_d < 200$, while magenta points relate to domains with $200 \leq n_d < 500$ and green points refer to domains where $n_d \geq 500$. The red line within each panel indicates the 95 per cent nominal coverage rate, whereas the black line shows the estimated regression line from a linear regression of the coverage rates on the skewness coefficients. It can be seen that the slope of this line is negative in each panel, implying that an increase of the skewness coefficient is negatively associated with the coverage rates. This finding is intuitive as the confidence intervals are constructed using a normal approximation, which is less appropriate if the underlying data exhibits a high degree of skewness. Moreover, we observe an impact of domain-specific sample size in case of the direct estimator. Here, the worst coverage rates occur with the proportional allocation in domains where $n_d < 500$. In contrast, the indirect GREG estimator can compensate for smaller sample sizes by borrowing strength but suffers as well from the skewness of the variable of interest.

6.1.3 Results for national estimates

The results for the estimation of the national mean of the labour costs are shown in Table 6.4. Since design-based or model-assisted estimators are typically used to produce national estimates, we focus on them. Regarding the quality of the point estimates, we note that both estimators do not exhibit relevant Monte-Carlo biases in any of the designs. With respect to the RRMSE of the point estimates, we see that the proportional allocation yields the worst results for both estimators. Furthermore, for the direct estimator, the Costa-type allocation as well as the equal allocation and the CRH procedure yield very similar results. If a GREG estimator is used, the equal allocation performs slightly worse than the Costa-type and CRH allocations. Clearly the best results for both estimators emerge under the optimal allocation. The results for the mean confidence interval length follow a similar pattern to the RRMSE for both estimators, with the proportional allocation as the worst and the optimal allocation as the best choice. In terms of the average coverage rate,

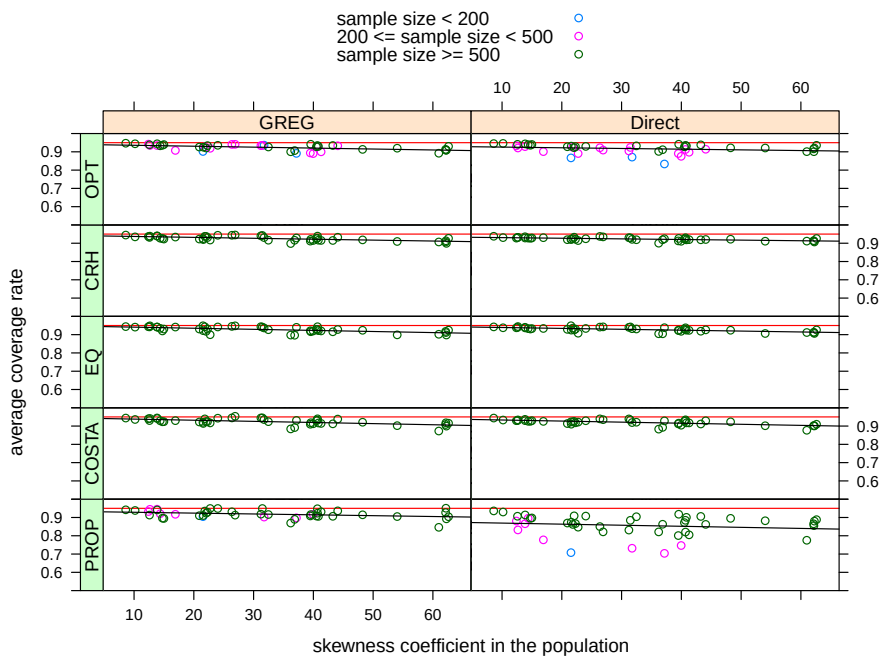

Figure 6.3: The relationship between average coverage rates and the skewness coefficient of small area estimators for the simulation study on Italian business data

we observe a slight undercoverage in most designs, except for the direct estimator under the optimal allocation. It should be further noted that the RRMSE of the direct estimator is slightly larger than the desired threshold of 1.5 per cent. This finding is probably due to the fact that the optimal sample sizes were rounded to the nearest integer, which may result in a small loss of the total sample size.

6.1.4 Summary

In this study, we examined the performance of various estimators under stratified designs for a business survey. Our variable of interest is severely skewed to the right, which is a common feature in many business surveys. Hence, before modelling the variable of interest, a non-linear transformation has to be applied. In the present case, we considered the log transformation for this purpose, as a rich body of estimation techniques is available for lognormal mixed models (cf. Section 3.4). While the ALLOG estimator based on the area level lognormal mixed model delivered good results in general, the EBP under the unit level lognormal mixed model performed rather poorly.

One explanation in this regard relates to the validity of the normality assumptions on the random effects and individual error terms. Therefore, we inspected QQ plots of the transformed residuals from the unit level lognormal mixed model for the first sample in each design. Figure 6.4 depicts the QQ plot for the Costa allocation with $c = 0.5$. The plot clearly indicates a violation of the normality assumptions, which are used to derive the EBP. Hence, the poor performance of the EBP is not fully surprising. One possible alternative is to consider a Box-Cox transformation of the response vector. We tried this option, but the optimal value of the transformation parameter λ was very close to zero, in which case the Box-Cox transformation reduces to the log transformation.

Table 6.4: Simulation results of the national mean estimates for the study on Italian business data

		Designs				
Measure	Estimator	PROP	COSTA	EQ	CRH	OPT
Relative Bias	GREG	-0.00012	-0.00001	-0.00011	-0.00005	0.00014
	Direct	-0.00033	-0.00007	-0.00016	0.00002	0.00016
RRMSE	GREG	0.01712	0.01296	0.01342	0.01296	0.00855
	Direct	0.02018	0.01507	0.01517	0.01517	0.00962
ACR	GREG	0.9436	0.9471	0.9429	0.9414	0.9487
	Direct	0.9453	0.9484	0.9429	0.9462	0.9511
Mean CI length	GREG	2634.41	1989.61	1969.55	1960.44	1309.77
	Direct	3105.05	2329.92	2254.57	2299.18	1481.06

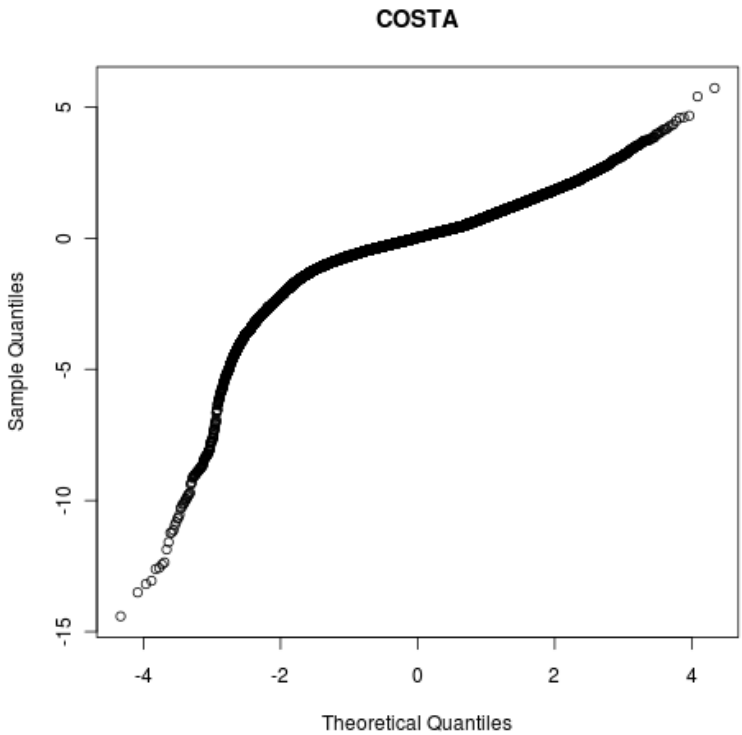

Figure 6.4: QQ plot of the transformed residuals for the Costa allocation

Thus, we conclude that the problems with the unit level EBP are not so much caused by a suboptimal transformation. A more likely reason for the poor performance of the EBP is that the access to strong predictive covariates is rather limited. Interestingly, the worst results for the unit level EBP all occurred in domains where the first digit of industry classification is a 8, which suggests that this branch behaves differently from the other sectors. Hence, we tried to include the first digit as a covariate, but it did not yield substantially different results.

Instead, if a model-based estimator is to be used, one should focus on area level models. In our study, the ALLOG predictor modelling the log of the direct estimates delivered good results. One explanation in this regard is that each direct estimate is computed by aggregating individual observations and thus the model assumptions may be easier to fulfil. Moreover, the domain-specific sample sizes were relatively large in this application such that the variance of the direct estimator was already small in many cases. In these situations, the ALLOG tends to behave like the direct estimator and is therefore protected against potential consequences of a misspecified model. Moreover, as the assisting model used in our study had some predictive power, the GREG estimator can realise variance reductions versus the direct estimator. In terms of the allocation procedure employed in the StrRS scheme, we saw that the CRH allocation can be very efficient for domain estimates when the stratum-specific standard deviations are known. This property does not only hold for the direct and the GREG estimators but also for the ALLOG predictor, which uses the direct estimates as an input. In terms of the confidence interval coverage rates, we observed some undercoverage for the direct, GREG and ALLOG estimators. We could explain this phenomenon the first two estimators by the skewness of the dependent variable in the underlying domain. This suggests to replace the quantiles from the normal distribution by those obtained from a suitable resampling approach, say a bootstrap or jackknife.

Moreover, we compared various allocation procedures of the total sample size to the strata. Here, a trade-off between efficient estimation on the national level and obtaining reliable domain estimates clearly emerged. If our focus is on the national level, then the optimal allocation is clearly the best choice. For the purpose of domain estimation, however, other procedures except the proportional allocation are clearly superior. Note that in our study the CRH allocation yields very good results for domain estimation and is less favourable for estimating the national mean. This finding is probably due the specific tolerances chosen in this application.

6.2 Domain estimation of employment characteristics

Labour force surveys (LFS) are carried out in order to gain information about employment characteristics of the participants (cf. Eurostat, 2001, p. 4). Typically, these surveys are designed to provide design-based estimates of aggregate statistics with a reliable precision. Besides these aggregate estimates, figures for subgroups defined by regional or socio-demographic partitions or both may be warranted as well. An often discussed example is the phenomenon of high unemployment rates amongst young people in the European Union, which recently attracted much attention (cf. Tse & Esposito, 2013). As already

pointed out in Chapter 2, providing reliable design-based estimates for these domains may be impossible if the sampling design does not explicitly include these subgroups, as the resulting sample sizes may be too small.

6.2.1 Setup

The Luxembourgian statistical office (STATEC) kindly provided data from the Census 2011 which also serves as a frame to draw the Luxembourgian LFS. The sampling design used is a simple random sampling of households, where $l = 8{,}000$ households are drawn without replacement. It should be noted that most labour market statistics such as the total of employed or unemployed people are computed for individuals. Hence, the survey design of the Luxembourgian LFS amounts to a single-stage cluster sampling with households as clusters. We repeated the procedure of drawing the sample $R = 10{,}000$ times in order to conduct a design-based simulation study.

The variable of interest is the employment status, which will be partitioned in three categories: employed, unemployed and economically inactive. As domains of interest we consider the cross-classification of the seven age classes with an indicator of the nationality (Luxembourgian: yes/no) and the gender, which yields $D = 28$ domains in total. In practice, also geographical information could be used to construct the domains of interest, but our data set did not include this information.

Since our dependent variable is nominally scaled, we consider models appropriate for this type. To capture the nominal structure we could use a multinomial logit model, briefly mentioned in Section 3.6. The estimation of a multinomial mixed model is very complex and was not pursued further. Alternatively, binary regression models employing the logit or probit transformation can be employed, where separate models are used for the different categories of the variable of interest. In addition to model-based estimators, we examined various model specifications for a GREG estimator and considered also a direct estimator.

The different estimators employed in this study are summarised in Table 6.5.

Table 6.5: Estimators for the simulation study on the Luxembourgian LFS

Abbreviation	Description
HT	The HT estimator of the domain total
GREG	Indirect GREG estimator, predictions are obtained from a linear fixed effects model
LGREG	Indirect GREG estimator, predictions are obtained from a logistic fixed effects model
MLGREG	Indirect GREG estimator, predictions are obtained from a logistic mixed effects model
MultGREG	Indirect GREG estimator, predictions are obtained from a multinomial logistic fixed effects model
BHF	EBLUP given by (3.2.7)
BINP	Naive predictor under a two-level logistic mixed effects model

6.2.2 Results for totals

The distribution of the relative biases for the different estimators over the domains for the employed labour force is given in Table 6.6. In this table the first column indicates the estimator while the columns *min* and *max* show the minimum (maximum) relative bias over all domains. The columns *median* and *mean* refer to the respective quantities of the relative biases and q_{25} and q_{75} denote the first (third) quartile of the distribution of the relative biases.

Table 6.6: Relative biases of the employed labour force in the simulation study on the Luxembourgian LFS

	min	q_{25}	median	mean	q_{75}	max
BHF	-1.369e-03	-6.066e-04	-2.649e-04	2.237e-04	1.072e-03	2.882e-03
BINP	-1.578e-01	-7.905e-04	3.405e-02	3.513e-02	8.039e-02	2.933e-01
GREG	-5.936e-04	-8.824e-05	5.688e-05	5.751e-04	1.003e-03	3.371e-03
LGREG	-1.052e-03	-1.457e-04	5.605e-05	1.548e-04	3.474e-04	1.984e-03
MLGREG	-1.106e-03	-1.801e-04	5.650e-05	1.388e-04	3.766e-04	1.887e-03
MultGREG	-1.081e-03	-1.473e-04	4.861e-05	1.161e-04	3.574e-04	1.650e-03
HT	-9.488e-04	-2.275e-04	1.193e-04	2.402e-04	4.469e-04	2.024e-03

It is immediately obvious from Table 6.6 that the design-based and model-assisted procedures do not suffer from Monte-Carlo biases. This property is not surprising, since these estimators are (approximately) design-unbiased by construction. Moreover, the BHF estimator, which is the EBLUP under a unit level linear mixed model, is unbiased. This indicates that a linear relationship between the mean of the dependent variable, which is binary in this case, and the mean of our auxiliary information is reasonable. The model-based estimator based on unit level logistic mixed models, BINP, is severely biased.

The results for the relative root mean squared errors of the employed labour force is given in Table 6.7.

Table 6.7: RRMSEs of the employed labour force in the simulation study on the Luxembourgian LFS

	min	q_{25}	median	mean	q_{75}	max
BHF	0.01252	0.02054	0.02672	0.03980	0.06226	0.08512
BINP	0.01068	0.04112	0.08032	0.08729	0.11190	0.29740
GREG	0.01361	0.02166	0.02678	0.04153	0.06863	0.08673
LGREG	0.01251	0.02084	0.02671	0.03870	0.05905	0.08373
MLGREG	0.01250	0.02055	0.02667	0.03851	0.05866	0.08314
MultGREG	0.01250	0.02085	0.02674	0.03864	0.05875	0.08393
HT	0.03496	0.04153	0.04402	0.05747	0.07368	0.10280

Table 6.7 shows that our assisting models have some predictive power as the model-assisted GREG procedures have a distinct advantage compared to the HT estimator. We further note that the linear GREG performs just a little worse than the other GREG

estimators which yield similar results. Also the BHF estimator performs very much alike these estimators and a little better than the linear GREG estimator. The RRMSE of the BINP is dominated by the squared bias which is substantial for this method.

The results for the relative biases and RRMSEs of the unemployed labour force are given in Tables 6.8 and 6.9.

Table 6.8: Relative biases of the unemployed labour force in the simulation study on the Luxembourgian LFS

	min	q_{25}	median	mean	q_{75}	max
BHF	-1.622e-01	-4.796e-02	-6.038e-03	3.435e-03	7.093e-02	2.369e-01
BINP	-2.954e-01	-3.206e-02	3.359e-02	5.265e-02	1.605e-01	4.712e-01
GREG	-3.952e-03	-1.051e-03	-4.718e-04	-3.380e-05	1.422e-03	4.473e-03
LGREG	-4.091e-03	-1.107e-03	-5.030e-04	-2.622e-05	1.271e-03	5.175e-03
MLGREG	-4.169e-03	-1.138e-03	-4.851e-04	-3.897e-05	1.271e-03	5.156e-03
MultGREG	-4.257e-03	-1.077e-03	-8.365e-04	-2.189e-04	1.037e-03	5.423e-03
HT	-3.788e-03	-1.105e-03	-1.861e-04	-3.505e-06	1.096e-03	5.312e-03

The results for the relative biases for design-based and model-assisted estimators are very similar to the ones for the estimation of the employed labour force, i.e. they are unbiased. The BHF, however, exhibits now non-negligible relative biases in some domains, which indicates problems with the model assumptions. Furthermore, the BINP is severely biased.

Table 6.9: RRMSEs of the unemployed labour force in the simulation study on the Luxembourgian LFS

	min	q_{25}	median	mean	q_{75}	max
BHF	0.09783	0.13250	0.15870	0.17610	0.20740	0.30960
BINP	0.08571	0.11440	0.17440	0.19460	0.24400	0.49420
GREG	0.12020	0.15840	0.19210	0.19880	0.22610	0.31580
LGREG	0.12010	0.15840	0.19210	0.19870	0.22550	0.31550
MLGREG	0.12010	0.15840	0.19210	0.19860	0.22530	0.31560
MultGREG	0.12000	0.15840	0.19200	0.19830	0.22380	0.31540
HT	0.12520	0.16330	0.19610	0.20270	0.23000	0.31900

Table 6.9 shows that there is not much difference between model-assisted estimators and the HT estimator for the estimation of the unemployed labour force. This highlights that the assisting models do not have substantial predictive power, unlike the prediction for the employed labour force. Moreover, the BHF estimator, despite being biased has a lower RRMSE over the entire distribution than all design-based or model-assisted procedures. This indicates a trade-off between bias and variance of the estimates.

To assess the validity of point and MSE / variance estimates at the same time, we plot the confidence interval coverage rates against the average confidence interval length. Since the point estimates of the BINP were not satisfactory for both categories, we did not consider

MSE estimation for this predictor. The behaviour of the other estimators is depicted in Figure 6.5.

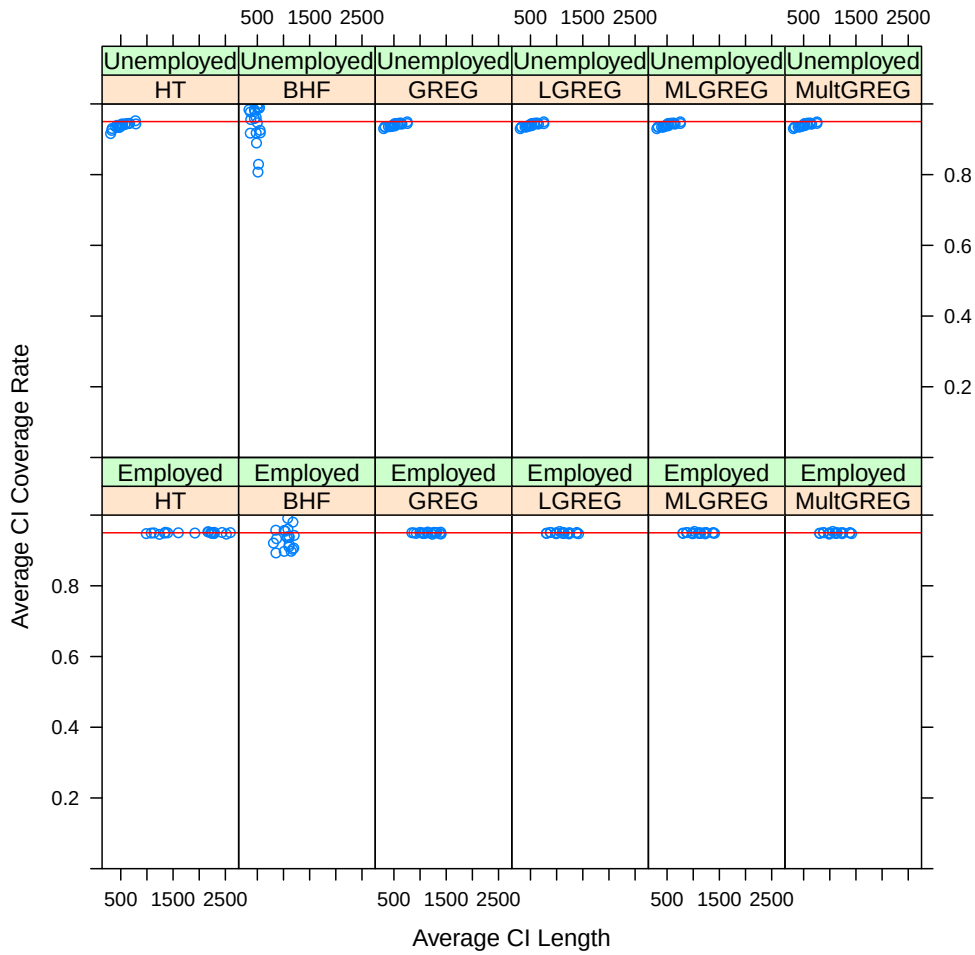

Figure 6.5: Confidence interval coverage rates of the totals in the simulation study on the Luxembourgian LFS

On the horizontal axis, we see the average confidence interval length and on the vertical axis the average coverage rate. The lower part of Figure 6.5 shows these quantities for the employed labour force in the domains, whereas the upper part relates to the unemployed labour force. In each panel, a horizontal red line is drawn, indicating the desired nominal coverage rate of 95%. It can be seen that for the prediction of the employed labour force the design-based and model-assisted strategies reach the nominal coverage rates in all domains. We may further note that the predictions from each of the model-assisted estimators are almost identical to the others. The predictive power from the assisting models leads to a higher precision as can be seen from the shorter confidence intervals compared to the HT estimator. These results agree with the ones obtained for the RRMSEs of the point estimates. The model-based BHF predictor has on average smaller intervals than the other estimators but its coverage oscillates between 90 and 99%, revealing that not all of the model-assumptions hold in this application.

For the prediction of the unemployed labour force, we hardly observe any differences between the HT estimator and the model-assisted procedures. For all these estimators

slight undercoverage occurs with the shortest prediction intervals. These short intervals are due to very small proportions of unemployed in some domains which translate into small variance estimates. The confidence interval lengths of the BHF are not much smaller than those of the other estimators. A drawback of the BHF predictor is that in some domains the coverage is as low as 80 % when a 95% coverage rate has been desired.

6.2.3 Results for unemployment rate

In addition to totals of the employed and the unemployed labour force, the unemployment rate is very important. The unemployment rate p_d in area d is defined as

$$p_d = \frac{\tau_{d,unem}}{\tau_{d,unem} + \tau_{d,emp}}, \quad d = 1, \ldots, D, \tag{6.2.1}$$

where $\tau_{d,unem}$ denotes the total of unemployed in domain d and $\tau_{d,emp}$ denotes the total of employed in domain d (cf. Molina et al., 2007). An estimate of the unemployment readily follows by replacing the true quantities $\tau_{d,i}$ with estimates $\widehat{\tau}_{d,i}$ where i refers to the categories *emp* and *unem*. Note that this leads to an estimate $\widehat{p}_d$ which is a ratio of random variables in both the numerator and the denominator. Hence, the estimates will suffer from a bias, which could be avoid if the denominator was fixed. This would be the case if both the number of all persons within each domain and the number of economically inactive persons within each domain were known. However, in our application, we consider these quantities as unknown.

To obtain an estimator of the variance of $\widehat{p}_d$, we make use of the Taylor linearisation. This technique is suitable since both the numerator and denominator are functions of totals. Applying the linearisation formula yields (cf. Lehtonen & Pahkinen, 2004, Section 5.3):

$$\mathrm{Var}(\widehat{p}_d) = \frac{\tau_{d,unem}^2 \mathrm{Var}(\widehat{\tau}_{d,emp}) + \tau_{d,emp}^2 \mathrm{Var}(\widehat{\tau}_{d,unem}) - 2\tau_{d,unem}\tau_{d,emp}\mathrm{Cov}(\tau_{d,unem}, \tau_{d,emp})}{(\tau_{d,unem} + \tau_{d,emp})^4}. \tag{6.2.2}$$

Whereas the variance estimates in (6.2.2) are easily obtained, the computation of the covariance term may be more complex. An exception is the direct estimator, where an analytical solution can be given. In the case of the model-assisted estimators we used the residual covariance. Alternatively, resampling methods such as the bootstrap or the jackknife could be employed. The relative biases and the RRMSEs are depicted in Figure 6.6.

It is evident from Figure 6.6 that the design-based and model-assisted estimators are unbiased, even though the target quantity is a ratio. The BINP and the BHF estimators both exhibit biases in many domains. With respect to the RRMSEs, we do not observe visible differences between the direct estimator and the model-assisted estimators. Furthermore, we can improve on these results by using the BHF estimator. However, this gain is not large. The unit level estimator BINP performs the best in terms of median RRMSE and minimum RRMSE, but suffers in some other domains. Hence, it yields the by far largest maximum RRMSE due to its large absolute relative biases in some domains.

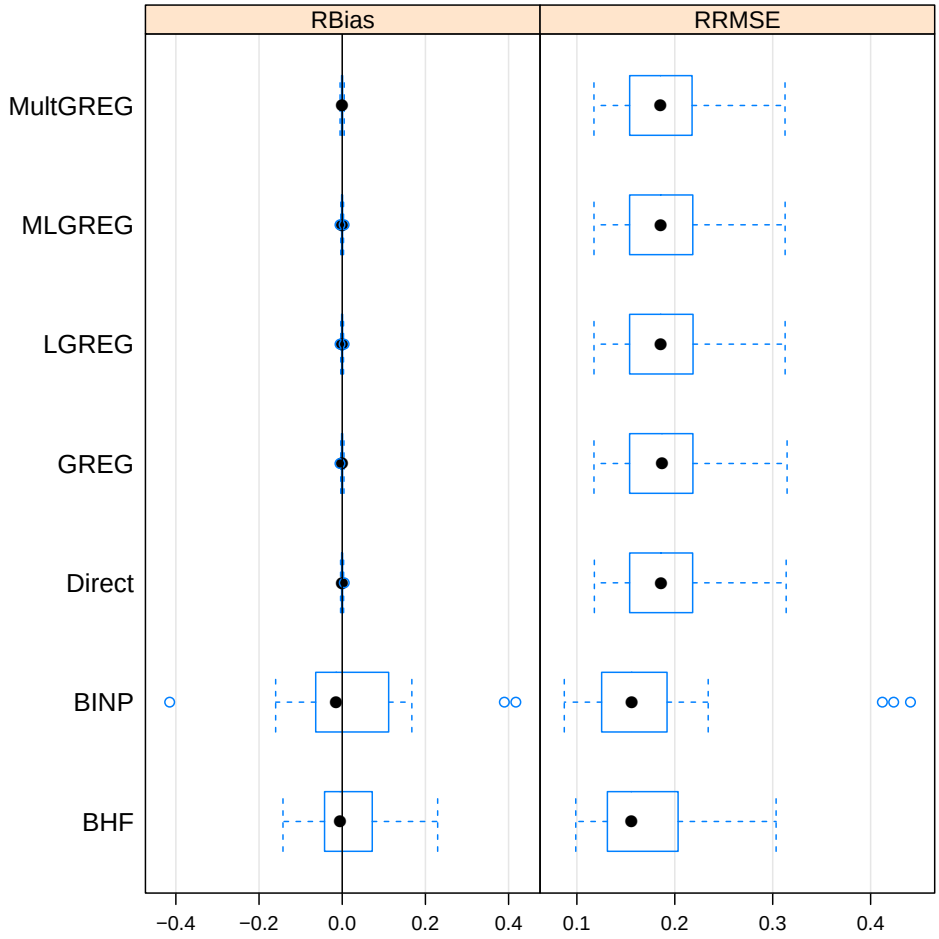

Figure 6.6: Relative biases and RRMSEs for the unemployment rates in the simulation study on the Luxembourgian LFS

To assess the inferential qualities of the predictors, we compared the average confidence interval coverage rates. We did not consider MSE estimation for the BHF and BINP estimators, which require the evaluation of the mean cross-product error between the number of employed and unemployed in each domain (cf. Molina et al., 2007). Table 6.10 gives the average confidence interval coverage rates for the different predictors. It can be seen that there are hardly any differences between the predictors. Moreover, slight undercoverage has to be acknowledged.

6.2.4 Summary

In this study, we examined several approaches to produce predictions for domain totals of the employed and the unemployed in the Luxembourgian LFS. To assess the performance, we conducted a design-based analysis, where the true sampling design was used. Regarding the prediction of the total of the employed labour force, model-assisted estimators yielded RRMSEs below 9 per cent for all domains. This finding is promising, since in practice more and stronger covariates may be available. Nonetheless, the results indicate the usefulness of incorporating auxiliary information at the estimation stage. Unfortunately, our auxiliary information is not rich enough to provide efficiency gains for

Table 6.10: Average coverage rates for unemployment rate in the simulation study on the Luxembourgian LFS

	min	q_{25}	median	mean	q_{75}	max
Direct	0.9298	0.9360	0.9388	0.9397	0.9441	0.9476
GREG	0.9288	0.9371	0.9389	0.9400	0.9432	0.9491
LGREG	0.9311	0.9360	0.9389	0.9399	0.9435	0.9497
MultGREG	0.9319	0.9361	0.9387	0.9401	0.9433	0.9495

the prediction of the total of the unemployed labour force. Hence, the results of model-assisted procedures are very similar to the HT estimator. In most cases, the public will care the most about estimates of the unemployment rate. In our simulation study, similar results were obtained for different variants of the GREG estimator and the direct estimator. This finding is due to the fact that the assisting model has little predictive power for the modelling of the unemployed labour force. Slightly better results in terms of the RRMSE emerged with the BHF estimator. However, they are accompanied by an increase of the bias. Moreover, in many cases, model analysts may have access to strong auxiliary information such as employment data from registers. In this case, using model-based estimators may yield much better results than using direct estimates as the findings of Pereira, Mendes, and Coelho (2013) indicate.

6.3 Small area estimation of poverty measures

In this section, we consider the estimation of the ARPR within a design-based environment. A detailed account on this study is given in the book chapter of Burgard, Münnich, and Zimmermann (2015). This application is chosen as it is very relevant for the poverty measurement in the European Union (EU). The analysis of poverty in the EU is based on the so-called Laeken indicators with the EU survey on income and living conditions (EU-SILC) as the most important data source. To allow for a realistic consideration, we choose the AMELIA data set that emerged as an important outcome of the AMELI project within the seventh framework programme of the European Commission. AMELIA is a fully synthetic data set based on structural relationships found in the EU-SILC surveys of the years 2005 and 2006. A detailed description of the AMELIA data is given in Alfons et al. (2011).

6.3.1 Sampling designs

As pointed out by Graf, Wenger, and Nedyalkova (2011), most countries participating in the EU-SILC survey use a one- or two-stage design to gather the sample information. Moreover, the designs are implemented separately within the participating countries. Thus, also the designs employed in the countries differ in general. To facilitate the implementation in our study, we assume that a common design is used in all survey countries.

Therefore, following Münnich and Burgard (2012), we decide to consider two-stage strat-

ified designs where the first stage represents the $D = 78$ small areas for which estimates are desired. On the second stage, we construct strata based on the realisations of the total disposable household income. This design is not feasible in real-life applications since the total disposable household income is a variable collected in the EU-SILC survey, but it helps to illustrate the maximum efficiency gains. The total disposable household income is chosen to construct the strata as it is related but not identical to the equivalised household income, which identifies the poor households.

In order to be more specific about the design impact of stratified sampling, we have to distinguish two separate effects. On the one hand, there is the stratification effect, leading to potential efficiency gains due to the construction of strata with a small within variance. On the other hand, there is the allocation effect, which is due to a specific choice of how to allocate the total sample size to the different strata. Both the stratification and the allocation are important decisions to be taken by the survey planner, which then along with the choice of the estimation technique contribute to the quality of the estimates.

In our design-based simulation study with $R = 10{,}000$ replications, we compare three different stratification strategies shown in Table 6.11.

Table 6.11: Stratification for the poverty analysis

Method	Description
rand	The households are randomly assigned to a stratum
sort	The households are assigned to a stratum based on the quantiles of their disposable income
raso	A mixture of the two above patterns

Source: Reprint from Table 1.1 in Burgard et al. (2015)

Our motivation to consider these choices of stratification is that they range from a rather weak to a very strong stratification effect. On the one extreme, the *sort* pattern reflects a favourable case for design-based estimation strategies. This is the case, since the stratification already accounts for much of the variation in a the total disposable household income. The *rand* pattern, however, refers to a situation that is less pleasant for design-based estimation, as the stratification is not related to the dependent variable at all. Regarding the allocation of the sample size to the strata, we compare the equal allocation, the proportional allocation, and the optimal allocation, which were introduced in Section 2.3.2. To illustrate the stratification effect, we also include SRS within the areas as a benchmark case. The resulting area-specific sample sizes are given in Table 6.12, where q_p denotes p-th quantile.

For the SRS case, we set the area-specific sample sizes to the corresponding ones for proportional allocation. It can be seen from Table 6.12 that only the equal allocation guarantees sample sizes above 30 units for all areas. Already the proportional allocation displays a large variation of the sample sizes between areas. While the largest area has a sample size of 385 units, only 10 units are drawn from the smallest area. This could be a major issue for design-based and model-assisted estimation strategies. Using the optimal allocation amounts to an even higher degree of variation in the sample sizes. Now the

Table 6.12: Sample sizes n_d for the simulation study on poverty measures

Design	prop	equal		opt	
Pattern			rand	raso	sort
min	10.00	75.00	10.00	10.00	11.00
$q_{0.25}$	56.25	75.00	51.00	52.00	47.00
$q_{0.5}$	75.00	75.00	69.50	70.50	65.00
mean	76.92	76.92	76.92	76.92	76.92
$q_{0.75}$	93.75	80.00	99.75	97.00	96.25
max	385.00	80.00	462.00	427.00	468.00

Source: Reprint from Table 1.2 in Burgard et al. (2015)

sample sizes differ by a factor of more than 40 in all stratification patterns. Note that the sample sizes under equal and proportional allocation do not vary with the stratification pattern used.

6.3.2 Estimators

Our study compares design-based and model-assisted estimators to their model-based counterparts. As a design-based estimator we consider the direct estimator. To realise potential efficiency gains from an assisting model, we also include the GREG and LGREG estimators introduced in Section 2.2, using fixed effects assisting models estimated for the whole sample. Especially the latter estimator seems suitable for the estimation of the ARPR, as the dependent variable is binary on the household level, where our modelling approach operates. We also tried assisting models which use a random intercept, but they performed almost identically to models without a random intercept and were thus not included in the further analyses. As discussed in Section 2.2, this is due to treating the domains as planned in our study, which implies that the estimated domain effects cancel from the predictions under a linear model. Regarding model-based estimators, we included the unit level EBP due to Molina and Rao (2010), which was presented in Section 3.3. In order to ease the computation, we exploit the fact that the ARPR is a separable and additive measure, which allows us to use numerical integration instead of the Monte-Carlo integration. Besides the unit level EBP, we include predictors based on area level models of type (3.2.17). These procedures are commonly applied by the US Bureau of the Census to produce poverty estimates in the SAIPE project under a Bayesian framework (cf. https://www.census.gov/did/www/saipe/). We prefer to stay within the frequentist paradigm, however, and consider the original Fay-Herriot predictor (3.2.19). Noting that the direct estimator is a proportion, we may alternatively employ variance-stabilising transformations which additionally yield a better approximation to a normal distribution. A frequently used approach is to use the arcsine-square-root transformation and to model $\theta_d^{\text{Direct}} = \sin^{-1}(\sqrt{\text{ARPR}_d^{\text{Direct}}})$ (cf. Jiang & Lahiri, 2006b, Section 2.1 and references therein). This leads to the following two-level model used by Jiang, Lahiri, Wan, and Wu (2001):

$$\begin{aligned} &\text{sampling model} \quad && \widehat{\theta}_d^{Dir}|\theta_d \overset{ind}{\sim} N(\theta_d, \psi_d), \quad d = 1, \ldots, D, \\ &\text{linking model} \quad && \theta_d \overset{ind}{\sim} N(\overline{\mathbf{X}}_d^T \boldsymbol{\beta}, \sigma_v^2), \quad d = 1, \ldots, D. \end{aligned} \tag{6.3.1}$$

As outlined by those authors, ψ_d, the variance of the transformed direct estimator is approximately equal to the design effect divided by the area sample size times four. A simple predictor of the ARPR is then obtained by applying the back-transformation to the EBLUP of θ_d as

$$\widehat{\mathrm{ARPR}}_d^{\mathrm{FHtrans}} = \sin(\widehat{\theta}_d^{\mathrm{FH}})^2, \quad d = 1, \ldots, D, \tag{6.3.2}$$

where $\widehat{\theta}_d^{\mathrm{FH}}$ is the EBP of θ_d under model (6.3.1) obtained by employing (3.2.19) on the transformed variable θ_d. As pointed out by Rao (2003, Section 7.1), the predictor (6.3.2) is not the EBP for ARPR_d, since the back-transformation is non-linear. But it can be argued that if the area-specific sample sizes are not to small, (6.3.2) yields a reasonable approximation to the EBP (cf. Jiang et al., 2001). The expression for the EBP under model (6.3.2) can be derived by noting that it amounts to the expectation of a function of a random variable, i.e:

$$\widehat{\mathrm{ARPR}}_d^{\mathrm{AEBP}} = \mathrm{E}\Big(\mathrm{ARPR}_d | \widehat{\theta}_d^{\mathrm{FH}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}_v^2\Big). \tag{6.3.3}$$

In (6.3.3), $\widehat{\theta}_d^{\mathrm{FH}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}_v^2$ are estimates obtained from fitting model (6.3.1) to the data. While a closed-form expression for the expectation in (6.3.3) cannot be given, it can be evaluated using numerical integration. To pursue this idea further, we use the fact that the estimated posterior density of the transformed variable is given by $N(\widehat{\theta}_d^{\mathrm{FH}}, \widehat{\gamma}_d \psi_d)$ (Rao, 2003, Section 9.2.1). Since the EBP is simply the expectation of a transformation of $\widehat{\theta}_d^{\mathrm{FH}}$, it can be computed as

$$\widehat{\mathrm{ARPR}}_d^{\mathrm{AEBP}} = \int_{-\infty}^{\infty} \sin(\nu)^2 \frac{1}{\sqrt{2\pi\widehat{\gamma}_d\psi_d}} \exp\left(-\frac{1}{2}\left(\frac{(\nu - \widehat{\theta}_d^{\mathrm{FH}})^2}{\widehat{\gamma}_d\psi_d}\right)\right) \mathrm{d}\nu, \quad d = 1, \ldots, D. \tag{6.3.4}$$

The differences between the EBP (6.3.4) and the naive predictor (6.3.2) are most pronounced in cases where the domain-specific sample sizes are small, the design effect is large or a combination of both occurs. In our simulations, however, we did not observe substantial differences between the two predictors and thus report results only for the predictor (6.3.2).

As an alternative strategy, we observe that $n_d \mathrm{ARPR}_d^{\mathrm{Dir}}$ is an integer and use a binomial distribution for the sampling model. Hence, we modify model (6.3.1) in the following manner:

$$\begin{aligned} &\text{sampling model} && n_d \widehat{\mathrm{ARPR}}_d^{\mathrm{Dir}} | \mathrm{ARPR}_d \overset{ind}{\sim} Bin(n_d, p_d) \\ &\text{linking model} && \mathrm{logit}(\mathrm{ARPR}_d) \overset{ind}{\sim} N(\overline{\mathbf{X}}_d^T \boldsymbol{\beta}, \sigma_v^2), \quad d = 1, \ldots, D. \end{aligned} \tag{6.3.5}$$

Model (6.3.5) is an extension of the model developed by Jiang and Lahiri (2001) for binary data. Thus, also the EBP under (6.3.5) can be obtained in a similar fashion to their procedure. One concern that commonly arises with logistic mixed models of the kind (6.3.5) is the estimation of $\widehat{\sigma}_v^2$, as the variance of the random effect is typically close

to zero. Thus, the issues discussed in Chapter 3 of this thesis apply as well. This may lead to problems for the integration routine, as the integrals in the numerator and the denominator of the EBP may be estimated to zero and thus the EBP may collapse. In our study, we frequently encountered the issue of the random effect variance estimated to be zero and thus disregard model (6.3.5) for the further analyses.

Following the arguments of Datta, Hall, and Mandal (2011) discussed in Section 3.5, we decide to use a deterministic linking model instead, i.e

$$\text{logit}(\text{ARPR}_d) = \overline{\mathbf{X}}_d^T \boldsymbol{\beta}, \quad d = 1, \ldots, D. \tag{6.3.6}$$

This amounts to a synthetic model, since the unobserved heterogeneity in ARPR_d not explained by the covariates $\overline{\mathbf{X}}_d$ is not accounted for. To account for the sampling design as well, we modelled the estimated population total instead of the sampling total. This leads to the following binomial synthetic area level model:

$$\begin{aligned} &\text{sampling model} && \widehat{N}_d \widehat{\text{ARPR}}_d^{\text{Dir}} | \text{ARPR}_d \overset{ind}{\sim} Bin(\widehat{N}_d, \text{ARPR}_d) \\ &\text{linking model} && \text{logit}(\text{ARPR}_d) = \overline{\mathbf{X}}_d^T \boldsymbol{\beta}, \quad d = 1, \ldots, D, \end{aligned} \tag{6.3.7}$$

where $\widehat{N}_d = \sum_{j=1}^{n_d} w_{dj}$ is the estimated population size in domain d. Note that for the designs in preceding section, $\widehat{N}_d = N_d$. We use the following predictor

$$\widehat{\text{ARPR}}_d^{\text{SynAL}} = \frac{1}{1 + \exp(-\overline{\mathbf{X}}_d^T \widehat{\boldsymbol{\beta}})}, \quad d = 1, \ldots, D, \tag{6.3.8}$$

where $\widehat{\boldsymbol{\beta}}$ is obtained from fitting the model (6.3.7). Since predictor (6.3.8) is synthetic, its variance can be easily estimated, e.g. using the jackknife, but estimating its MSE is not straightforward. We follow the strategy of Marker (1995), which is discussed in Rao (2003, Section 4.2.4). This strategy assumes that squared design bias in each domain is close to the average squared bias, which in turn can be estimated from the average MSE minus the average variance. The MSE of (6.3.8) follows from applying the MSE identity, i.e. $\text{MSE} = \text{Var} + \text{Bias}^2$. Table 6.13 lists the predictors used in our study.

Regarding the choice of the covariates, a stepwise backward elimination procedure based on the marginal AIC (3.5.1) for a fixed effects logit model is used. This approach leads to selection the following covariates: the number of retired person in household (RB210old), the number of employed persons in household (RB210work), a dummy variable indicating whether any unemployed person lives in the household (RB210unem), the household type (HHT), the taxes on income (HY140C), the degree of urbanisation (DOU) and the region (REG). In addition to these variables, we also include the stratum membership as a covariate for the StrRS designs when unit level models are considered, as the stratification is related to the variable of interest. Note that with area level models, we can not include the stratum membership, as this would yield perfect multicollinearity due to the stratification within areas.

Table 6.13: Estimators used in the simulation study on poverty measures

Abbreviation	Predictor
DIR	Direct estimator
GREG	GREG estimator under a linear fixed effects assisting model
LGREG	GREG estimator under a logistic fixed effects assisting model
MR	EBP based on nested error regression model due to Molina and Rao (2010)
FH	EBP based on area level model without transformations
FHtrans	Naive predictor (6.3.2) based on area level model with arc-sine square-root transformation
SynAL	Predictor (6.3.8) based on synthetic binomial area level model

6.3.3 Results on point estimates

Figure 6.7 depicts the relative biases of the estimation strategies for the different designs. The green box above each panel indicates the stratification pattern used, while the rose box denotes the allocation. The panels are ordered such that the stratification effect increases from the top to the bottom. Starting with the second column, the ordering from left to right is such that we gradually move from equal to unequal domain-specific sample sizes. Note that the placement of the SRS design in this figure is arbitrarily chosen, as we wanted to include it in the same figure.

We see from Figure 6.7 that the design-based and model-assisted procedures DIR, GREG and LGREG are unbiased for all designs, as they should be. It may be further noted that the SynAL and the FHtrans may encounter some biases in some areas and designs, but they do not exhibit systematic biases in either design. For the untransformed FH predictor, however, we observe a tendency to underestimate the true value, especially if the optimal allocation is used. Finally, the unit level model-based MR predictor is severely positively biased in all designs. For the optimal allocation, we may further note that the bias of the MR increases from the top to the bottom.

The results for the RRMSEs are given in Figure 6.8. We note that the design-based and model-assisted procedures suffer from high variances of the point estimates, which lead to unacceptably large RRMSEs in some domains. By a careful selection of the design, however, it is possible to reduce the RRMSE of the DIR, GREG and LGREG estimators substantially. Part of this reduction can be attributed to the sample size, as the equal allocation yields the smallest maximum RRMSE for all stratification patterns. Moreover, also the stratification effect plays as major role, since the RRMSE is smallest for the equal allocation when the sort pattern is used. Further, it should be emphasised that this effect is most pronounced for the direct estimator, since its efficiency gains are solely due to the sampling design as it cannot use additional information at the estimation stage. Note that these results do not suggest the existence of a free lunch opportunity, as the better performance of direct and GREG-type estimators in equal allocation in most domains is at the expense of efficient estimation in a few domains, where the proportional allocation

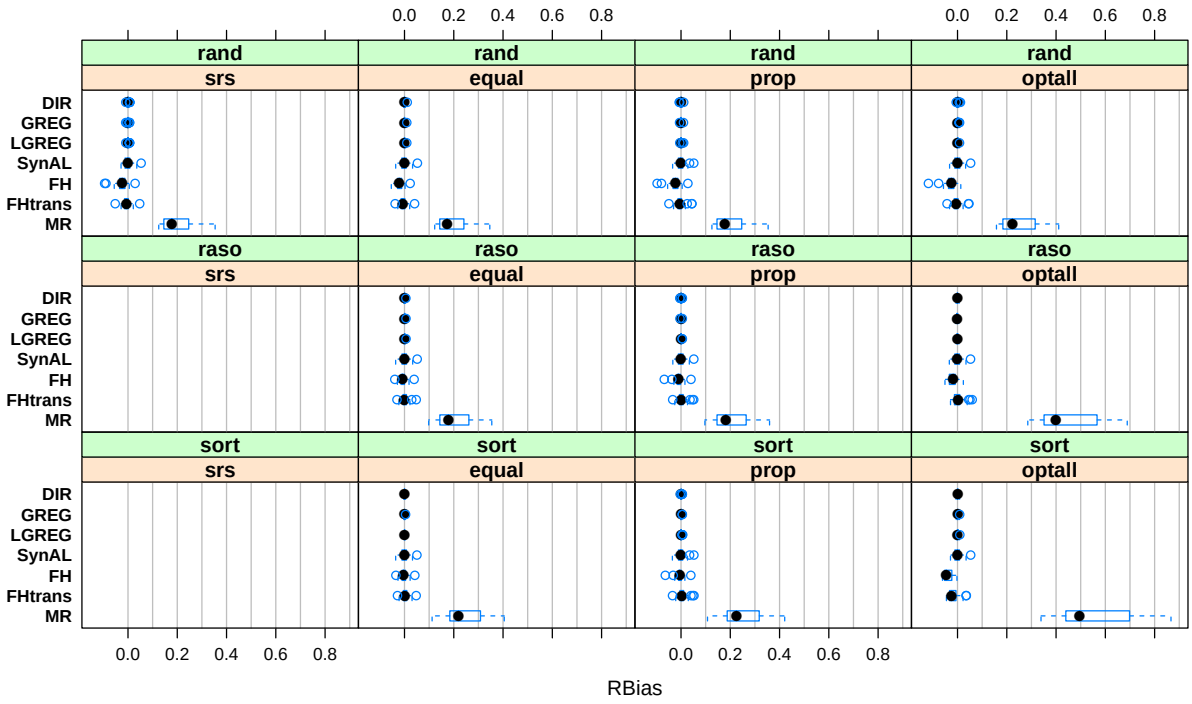

Figure 6.7: Relative biases in the simulation study on poverty measures

Source: Adapted from Figure 1.1 in Burgard et al. (2015)

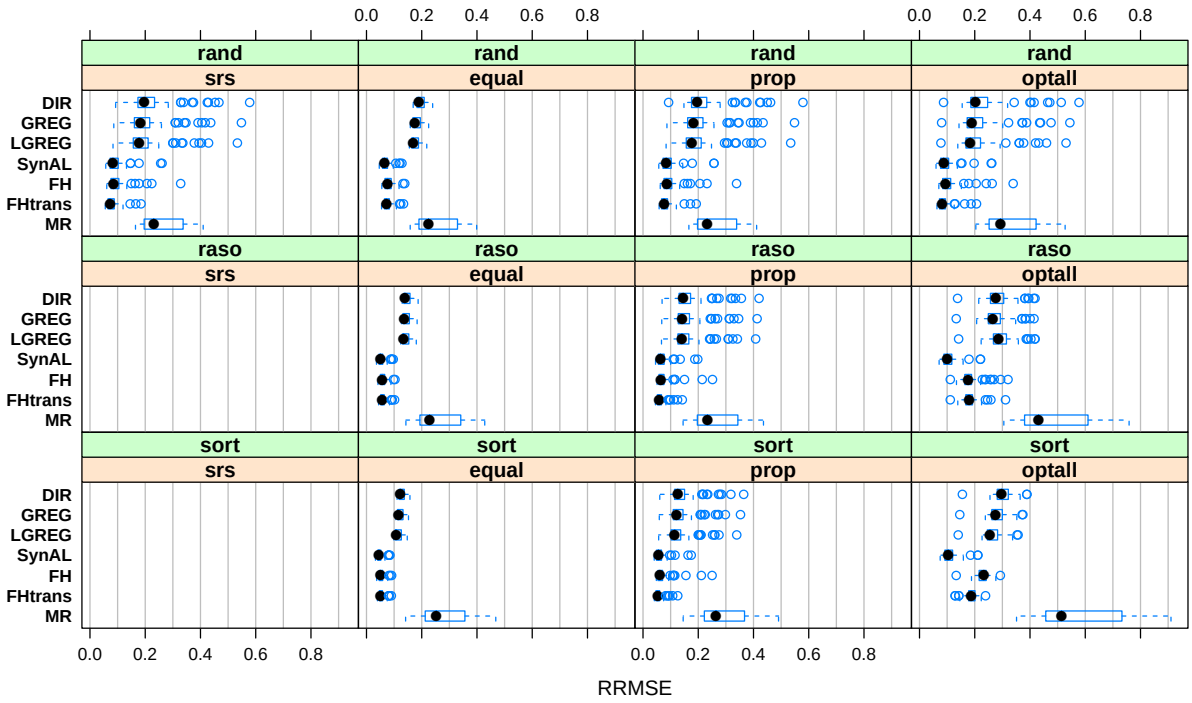

Figure 6.8: RRMSEs in the simulation study on poverty measures

Source: Adapted from Figure 1.2 in Burgard et al. (2015)

yields smaller RRMSEs. This is due to the fact that the maximal sample size over all areas is much larger under the proportional allocation as opposed to the equal allocation, which is visible in the last row of Table 6.12.

For the proportional allocation the gains due to the stratification are obvious. Whereas under the *rand* stratification the maximum RRMSE of the direct estimator is close to 60 percent, it is below 40 percent using the *sort* scheme. Perhaps even more importantly, we may reduce the median RRMSE from roughly 20 percent to only 12 percent by moving from the least to the most efficient stratification pattern. In general, the results obtained by SRS are very similar to the ones in proportional allocation with random assignment to the strata.

The results under the optimal allocation deviate substantially from the patterns observed for the other StrRS designs. Now, the minimum RRMSE is obtained under the *rand* pattern and it deteriorates for *raso* and also *sort* patterns. This is due to the fact that under the optimal allocation the highest sample sizes are realised in strata where the total disposable household income varies the most. Since this variable is right-skewed, those strata comprise the households with the highest total disposable incomes. However, typically only very few households with the high total disposable incomes are below the poverty threshold. Thus, the optimal allocation leads to high sample sizes in strata with hardly any variation of the variable of interest. Hence, the allocation which is optimal for the estimation of the total disposable income on the national level is not suitable for the estimation of the ARPR.

The choice of the assisting model for GREG-type estimators does not lead to very different results. Nonetheless, it should be noted that a logistic model yields slightly smaller RRMSEs when the strata are randomly assigned. This effect vanishes, however, for the equal allocation using a stronger stratification, e.g. the *raso* or *sort* scheme. Surprisingly, the results with optimal allocation and *raso* scheme differ altogether. In this design, the linear GREG performs better than its logistic counterpart. This issue will be explored in more detail in the following section.

With regards to performance of model-based predictors using area level information, we see that the equal allocation yields the best results in general. Furthermore, a strong stratification has fortunate consequences in this case, as the RRMSEs under the *sort* scheme are the smallest for these procedures. Interestingly, the SynAL predictor (6.3.8) using a deterministic logit-link performs better than both the FH and FHtrans under equal allocation. In the case of an optimal allocation, this effect is even more pronounced, as the RRMSE of the SynAL is now substantially lower than the corresponding values under both other procedures in *sort* and *raso* schemes. If, the sample size is allocated proportional to the population size, we note that the area level model with arc-sine square-root transformation performs better than the SynAL predictor. A comparison of the FHtrans and FH procedures indicates that the transformation leads to better results for most cases.

Finally, the RRMSE of the MR predictor is dominated by the squared bias component. We see that the RRMSE increases with the stratification effect and is worst for optimal allocation schemes.

6.3.4 The role of the assisting model

The results of the preceding section indicate that there are small but persistent gains if the assisting model uses a logistic rather than a linear model specification for most designs. This holds for all designs but the optimal allocation with the *raso* stratification. Now the question why the general relationship breaks down under this particular setting arises. To explain why this is the case, it will be helpful to think about the circumstances in which a GREG-type estimator is going to perform well. As mentioned earlier, the GREG estimator can be written in the following manner:

$$\widehat{\mathrm{ARPR}}_d^{\mathrm{GREG}} = N_d^{-1}\Big(\sum_{j\in \mathrm{U}_d} \widehat{y}_{dj} + \sum_{j\in \mathrm{S}_d} w_{dj}(\mathbb{I}(y_{dj} < z) - \widehat{y}_{dj})\Big). \tag{6.3.9}$$

The expression within brackets comprises two parts: the sum of the predictions for all units in the population within a particular domain and the sum of weighted residuals in this domain from the sample. If the second term in brackets of (6.3.9) is equal to zero, this implies that a bias correction is not necessary. To achieve a low bias of the model-based part $\sum_{j\in \mathrm{U}_d} \widehat{y}_{dj}$, one might want to include many covariates. However, this is generally accompanied by an increase in the variance of predictions, so that this does not solve the bias-variance trade-off issue. As a measure of the bias of the model-based predictions we consider the bias adjustment ratio

$$\mathrm{BAR}_d = \frac{\sum_{j\in \mathrm{S}_d} w_{dj}(\mathbb{I}(y_{dj} < z) - \widehat{y}_{dj})}{\sum_{j\in \mathrm{U}_d} \widehat{y}_{dj}}. \tag{6.3.10}$$

The larger the absolute value of (6.3.10), the larger the necessary adjustment to the model-based predictions. Note that the bias adjustment will be zero by construction in some circumstances under a direct GREG estimator (cf. Section 2.2 of this thesis). Thus, we want to study whether the differences between the linear GREG and the logistic GREG observed in the previous section are somehow related to the bias adjustment ratio (6.3.10). In order to do so, we display the averages of absolute values of this ratio in Tables 6.14 to 6.17. It can be seen from these tables that under all designs, the amount of bias correction is less for the LGREG compared to the GREG with the exception of optimal allocation and raso pattern given in Table 6.17. Furthermore, we note that the amount of bias correction is largest in this case.

	GREG	LGREG
Min.	0.156	0.152
1st Qu.	0.158	0.154
Median	0.161	0.157
Mean	0.163	0.158
3rd Qu.	0.167	0.162
Max.	0.173	0.168

Table 6.14: Bias adjustment ratios (SRS) in the simulation study on poverty measures

	rand		raso		sort	
	GREG	LGREG	GREG	LGREG	GREG	LGREG
Min.	0.140	0.135	0.111	0.109	0.094	0.087
1st Qu.	0.141	0.136	0.112	0.110	0.095	0.088
Median	0.142	0.137	0.113	0.111	0.096	0.089
Mean	0.142	0.137	0.113	0.111	0.096	0.089
3rd Qu.	0.143	0.137	0.114	0.112	0.096	0.090
Max.	0.144	0.138	0.115	0.112	0.097	0.090

Table 6.15: Bias adjustment ratios (StrRS equal) in the simulation study on poverty measures

	rand		raso		sort	
	GREG	LGREG	GREG	LGREG	GREG	LGREG
Min.	0.156	0.152	0.120	0.119	0.105	0.098
1st Qu.	0.158	0.153	0.122	0.120	0.106	0.099
Median	0.161	0.157	0.124	0.123	0.108	0.102
Mean	0.162	0.158	0.125	0.123	0.108	0.102
3rd Qu.	0.166	0.162	0.128	0.126	0.111	0.104
Max.	0.172	0.167	0.132	0.130	0.115	0.108

Table 6.16: Bias adjustment ratios (StrRS prop) in the simulation study on poverty measures

6.3.5 Results on precision estimates

The confidence interval coverage rates are plotted against the average confidence interval lengths in Figure 6.9. The confidence intervals are computed using a significance level of $\alpha = 0.05$, leading to a nominal coverage rate of 95 per cent. This nominal rate is highlighted by the horizontal red line in each panel. The green crosses, pink triangles, and blue circles represent the *rand*, *raso*, and *sort* stratification respectively. Owing the poor performance of the MR in terms of the point estimates, we do not perform MSE estimation for this predictor and hence can not report any results regarding the confidence intervals.

The top row presents the results under SRS within areas. The nominal coverage rate is met by the design-based and model-assisted procedures, but in some of the areas the average length is high. In this regard, the model-based estimators are the better choice, as their confidence intervals are shorter. Besides, the FH and the FHTrans estimator both overshoot the nominal coverage rate in most areas, while still delivering shorter confidence intervals. In contrast, the SynAL does not reach the 95% in any area, which is a sign that its MSE estimates are negatively biased as the point estimates do not exhibit biases. This finding is confirmed by the inspection of the relative biases of the MSE estimates, which is not reported here.

For the equal allocation in the second row, the three stratifications show similar results in terms of coverage rates for all estimators at hand. Compared to the SRS design, the shorter intervals for most of the domains are striking. Furthermore, the stratification

	rand		raso		sort	
	GREG	LGREG	GREG	LGREG	GREG	LGREG
Min.	0.166	0.162	0.214	0.233	0.223	0.207
1st Qu.	0.168	0.164	0.216	0.235	0.225	0.209
Median	0.171	0.167	0.218	0.236	0.227	0.211
Mean	0.172	0.168	0.218	0.236	0.227	0.210
3rd Qu.	0.177	0.172	0.219	0.237	0.228	0.211
Max.	0.182	0.177	0.221	0.238	0.230	0.213

Table 6.17: Bias adjustment ratios (StrRS opt) in the simulation study on poverty measures

effect yields a clear ordering, such that the *sort* stratification yields the shortest and the *rand* stratification the longest intervals. Again, all estimators but the SynAL meet the 95% nominal coverage rate.

Under the proportional allocation, the confidence interval lengths are much more dispersed than with the equal allocation, especially with design-based and model-assisted estimators. This is due to the varying sample sizes in the domain. The same holds for the model-based estimators as well, but the increases vis a vis the equal allocation is less pronounced. Furthermore, an undercoverage for design-based estimators can be found in some domains.

When using the optimal allocation, these effects are even stronger. As the optimality is computed on the total disposable household income, strata with highly variable total disposable household incomes have a high sample size and vice versa. With a binary target variable, this may result in some strata with a very high sample size, but no variation of the dependent variable at all. This issue mainly arises with the *sort* pattern. It should be further noted that for very small direct variance estimates, the FH predictor converges to the direct estimator. Hence, the shortcomings of the direct estimator directly affect the FH estimator.

6.3.6 Summary

This study investigated the use of small area estimation techniques to poverty measurement in a realistic design-based environment. The results indicate that model-based estimators using area level models may be preferred to their counterparts based on unit level information, when the assumptions underlying the unit level model are not met. In our study, the advantage of the area level model is not due to an informative sampling mechanism, as the design information was already used in the unit level model. Rather the aggregation behind the direct estimates tends to help the modelling. A comparison of the different modelling options for area level models indicates that the arc-sine square-root transformation yields better results in our study than an untransformed estimator. A further benefit of using the arc-sine square-root transformation versus the binomial model is that the model fitting is less sophisticated. Regarding the allocation procedure used in StrRS, our study indicates some of the pitfalls that can arise when optimising the design. One conclusion is that the optimal allocation should be used with care, as a slight change in the variable of interest can lead to results which are much worse than very simple allocation procedures.

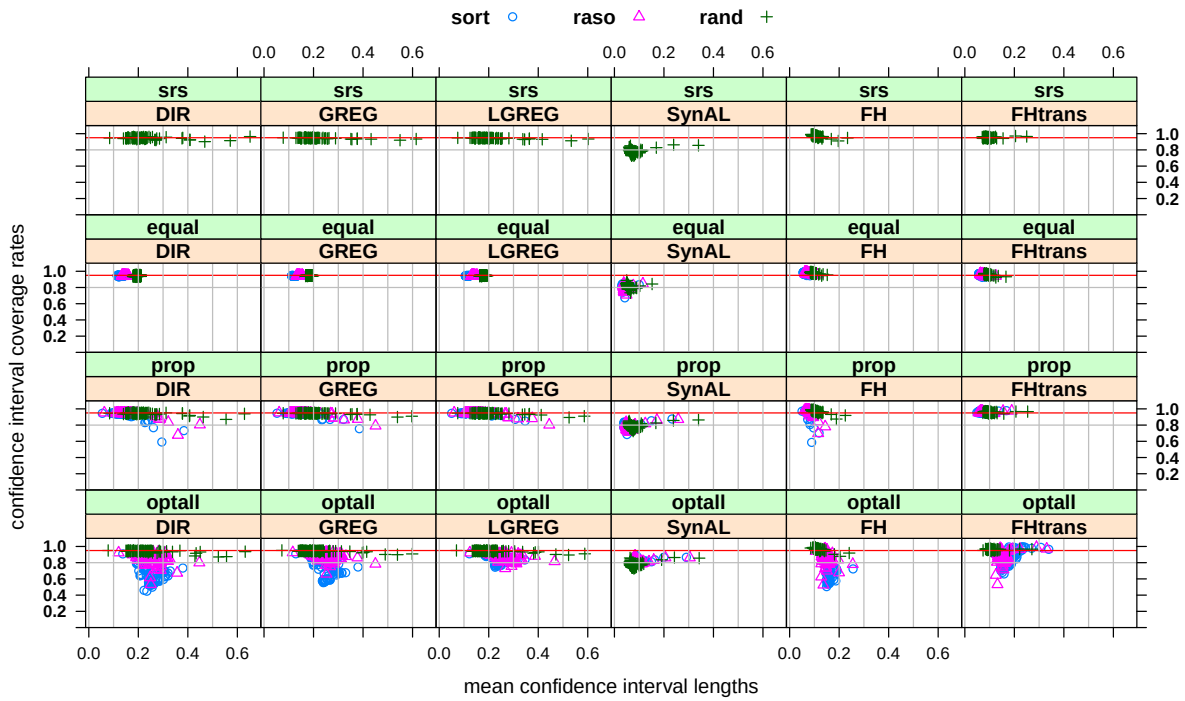

Figure 6.9: Confidence interval coverage rates in the simulation study on poverty measures
Source: Adapted from Figure 1.5 in Burgard et al. (2015)

Chapter 7

Conclusion and outlook

In this work we focussed on two complementary strategies to deal with the trade-off between design optimisation for aggregate statistics and the capability to employ model-based strategies for small area estimation. The first approach due to Verret et al. (2015), described in detail in Chapter 4, takes the design as given and corrects for a potentially informative sampling mechanism by including a function of the selection probabilities in the model. We extended this approach to the lognormal mixed model and derived the EBP under the augmented model. Moreover, a procedure to select the augmenting variable has been proposed in Section 4.4, which does not only account for the selection bias but also assesses the impact of the augmentation on the model assumptions. The results of model-based simulation studies under two different informative sampling mechanisms clearly showed that the proposed method corrects for non-ignorable designs. Moreover, with a high degree of informativeness, the augmenting variable led to a strong increase in the predictive accuracy such that very precise point estimates were obtained. Besides, the importance of checking the model assumptions after augmenting the model was revealed, as a violation of the normality assumption caused biases of the point and MSE estimates.

The second approach we considered to deal with the above mentioned trade-off is to construct a design which is well suited for the estimation of aggregate figures using design-based methods, but does not interfere with a model used for small area estimation. We proposed to draw a sample by means of a SRS of clusters that were constructed in an antithetic fashion. This leads to a non-informative sampling mechanism, provided the higher-level sampling is ignorable as well. Moreover, we established conditions in Section 5.2 under which our strategy is preferable to SRS for a number of underlying data generating models. The simulation studies reported in Section 5.3 further indicated benefits of this approach when using a model-assisted estimator under a misspecified model and for model-based small area estimation of the ARPR with estimated poverty thresholds. Moreover, ideas to adapt the proposed strategy for the case of more than one size variable have been proposed in Section 5.4.

Thus, this work provides substantial findings of practical interest when a survey planner wants to implement small area estimation in a design-based environment. The key steps can be summarised as follows: Firstly, following Marker (2001), important domains should be already taken into account at the design stage, such that design-based and model-assisted estimation procedures can provide reliable estimates. As this cannot eliminate the need for model-based estimation in small, unplanned domains altogether, care should be

taken not to impede modelling options after all. Ideally, this includes collecting variables known from a register which are related to the variables of interest. This is very important as the availability of strong auxiliary information is a prerequisite for building predictive unit level models. After collecting and preprocessing the sample, the next step is to identify a suitable model. For this purpose, the model selection tools and diagnostics reviewed in Section 3.5 can be used. If a reasonable unit level model can be found, it is important to check whether the sampling mechanism is informative for this model. If this is the case, an augmented model should be used, where the choice of the augmenting variable can be made based on model selection and diagnostics as well. Note that it might be possible that no suitable unit level model can be found after all. This can be due to scarce register information, an insufficient number of variables collected for modelling purposes or a combination of both. If this problem occurs, applying area level models can be a remedy as suggested by the applications in Chapter 6. Of course, the successful use of area level models requires the steps of identifying a suitable model and assessing whether the underlying assumptions can be maintained. An advantage of area level models is that the issue of informative sampling is less severe than for unit level models. This is due to the fact that in an area level model design-based direct estimates are modelled using register information. Hence, the sampling mechanism within areas poses no further problems. Note, however, that informative sampling may still arise due to a non-ignorable selection of areas (cf. Rao, 2003, Section 1.4). Thus, it is important to check for a potential informativeness of the area selection. Another modelling option is to relax the parametric assumptions and to employ a non-parametric approach instead (cf. Opsomer, Claeskens, Ranalli, Kauermann, & Breidt, 2008).

Further issues to be dealt with include the treatment of non-response and whether the domain estimates should be benchmarked against aggregate estimates. While the problem of benchmarking model-based estimators to the national total has been solved under a linear mixed model by You and Rao (2002a), the issue may stimulate further research for other models such as the lognormal mixed model. A drawback of simple solutions such as a ratio adjustment of the EBP under the augmented model is that a convenient solution regarding the estimation of the MSE has yet to be found.

Regarding the issue of non-response, it has been pointed out in Section 4.1.1 that erroneous inferences may result from the response behaviour as well. Note that the sampling and population models in a given area will coincide under non-response, if conditional on the covariates and the sample membership, the variable of interest and the response indicator are independent. This statement is similar to the requirement for a non-informative sampling mechanism given in (4.1.6). Hence, in principle, the same methods as those applicable to deal with informative sampling can be used. While it is conceptually straightforward to include all variables determining the response process in the model, this approach suffers from the same drawbacks as including all the design variables into the model. Moreover, resorting to augmented modelling may be more complicated in this case. Thus, we could look for a sufficient statistic which contains all the information about the response behaviour, such as the propensity score. However, there is a fundamental difference between the propensity score and the selection probabilities in that the latter are known from the sampling design, whereas the former needs to be estimated under a model (Pfeffermann, 2011, Section 2.1). Hence, we would include additional model uncertainty by employing an estimator under a model augmented by the propensity score. Thus, the validity of such an approach critically depends on the correct specification of the model for the propensity score. A solution applicable when the parameters of interest

are counts of a categorical variable in small areas is the procedure developed by Zhang (2009). He proposed to model the response probabilities by a logistic mixed model in an extension of the generalised structure preserving estimation approach due to Zhang and Chambers (2004).

Appendix A

Simulation studies for small area estimation

In many papers on small area estimation, simulation studies are used to evaluate the performance of estimators in a variety of situations. This approach allows to give recommendations in situations even in cases where analytical answers can not be obtained easily. In Section A.1, different simulation strategies relevant for small area estimation will be introduced. Section A.2 presents various measures, which could be used in a simulation study to assess the quality of the estimates.

A.1 Simulation frameworks

It was claimed in previous chapters that two competing modes of inference are of crucial importance for small domain estimation: the design-based and the model-based approach. Hence, we expect two frameworks for conducting simulations, one reflecting design-based simulation studies and another one related to model-based settings. Indeed, many papers report results from these two types of simulation approaches (e.g. Molina & Rao, 2010, Berg and Chandra (2014), Torabi and Rao (2008)). It should be noted, however, that both groups of simulation approaches experience a lot of within-group heterogeneity such that a further distinction seems worthwhile. In the group of model-based simulation studies, there are differences with respect to the treatment of the random effect. The unconditional expectation of the random effect is 0, but using this as the benchmark case is equivalent to stating that the true random effect is 0, which is hardly a reasonable assumption for small area estimation. Furthermore, design-based studies comprise settings in which the fixed finite population is taken as a realisation from a model and others in which there is no such data-generating process.

In the model-based framework, the observed values y_k are realisations from a superpopulation model, which specifies how certain input factors such as covariates and the domain-membership translate into the variable of interest (cf. Chambers & Skinner, 2003b, Section 1.3). Thus, a common feature in model-based simulations is that in all R Monte-Carlo replications realisations are drawn from the model. Hence, the randomisation distribution used for inferences is with respect to the model. It should be noted that in a model-based setting, there is no need to generate a finite population, but the samples can be drawn

directly from the assumed model. We will refer to such a setting, where the sample is drawn directly from the model as a purely model-based simulation. Obviously, if this approach is taken, it is very difficult to include a survey design. Moreover, we need to have true values against which the quality of the estimates can be assessed. As the generation of finite population quantities is circumvented, the inference will be directly based on characteristics of the superpopulation model. Thus, if the aim is to estimate the small area means

$$\mu_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}, \quad d = 1, \ldots, D, \tag{A.1.1}$$

we cannot compare the estimates directly to (A.1.1), as y_{dj} was generated for sampled elements only. Instead, the expected value under the model has to be used. For the case of frequently used linear mixed models this raises the question whether the expected value should be conditional on the realised random effect. If this is the case, the small area means in the r-th replication can be approximated as

$$\mu_d^{(r)} = \overline{\mathbf{X}}_d^T \boldsymbol{\beta} + v_d^{(r)}, \quad d = 1, \ldots, D. \tag{A.1.2}$$

An alternative is to draw a finite population in each Monte-Carlo replication, from which one sample is drawn. This strategy has been called a "design-model" approach by Verret et al. (2015). The above process will be repeated R-times, such that we have R different populations as realisations from the model and one sample for each of these populations. This procedure allows to evaluate each set of estimates derived from a particular sample against the values for the population from which that sample was drawn. If a mixed model is used, the random effect for a given population in a given area will be different from zero. Hence, this approach easily permits assessing the quality of the predicted random effects, which may not play a major role in the purely model-based study. Moreover, creating a finite population in each run enables to consider the impact of a sampling design on the estimator as well. However, it should be kept in mind that for model-based estimators, the randomness is due to the statistical model which was used to generate the y-values. As an implication, in principle, the process of drawing the sample can be deterministic for model-based estimators and will not lead to erroneous inferences, provided the model which holds for the sample holds also for the population. Note that the construction of finite population simplifies the introduction of design-based estimation strategies. Obviously, this requires the samples to be drawn by means of a probability sample.

We should note that in some model-based studies the values of the auxiliary information are generated only once and then held fixed throughout the study. This approach is valid as there is typically no assumption requiring the covariates to be random variables as well. Hence, only the random effects and error terms need to be generated in each run (Burgard, 2013).

Among the advantages of any model-based simulation study is that the behaviour of the data is entirely under control of the researcher. Hence, it can be ruled out that the results are due to the impact of omitted variables, which could not be mitigated. This

enables to attribute the measured results to a particular cause, which comes close to the experimental ideal. Moreover, a violation of the model's assumptions can be considered explicitly to serve as a robustness check against potential misspecification. Nonetheless, the external validity of the findings could be challenged, because real data rarely follow models, as the famous quote due to (Box & Draper, 1987, p. 424) "Essentially, all models are wrong, but some are useful" suggests. Thus to defend the insights obtained from a model-based simulation study, special emphasis should be laid on the study's applicability to real world problems.

Another simulation strategy is to draw a finite population once as a realisation of a superpopulation model and to keep it fixed throughout the study. Then in each Monte-Carlo replication, a sample is drawn according to a prespecified sampling design. In this case, the inference cannot be based on the model randomisation distribution, as we only have one realisation of the model. Hence, design-based inference is desired, which requires the sample to be drawn by means of a probability sample as defined in Section 2.1. Hence, the target quantities are unambiguously defined by the finite population at hand. We refer to such a setting as a quasi-design-based simulation study. This type of study is frequently used when model-based estimators are assessed in a design-based framework. Since the underlying model is known, issues such as model misspecification are avoided and the impact of sampling design can be evaluated.

The quasi-design-based approach might be subject of criticism, since the underlying data generating process is unknown in practice. Hence, it might be preferable to use an actual sampling frame / register which is then used to draw the sample. This process is very close to the survey practice and the situation commonly encountered in Official Statistics. A point of departure is that within a simulation study, we do not collect additional information. A typical solution is to include the variables of interest already in the register. To do so, the registers are enhanced by incorporating additional variables, which may refer to different scenarios. Examples include the AMELIA data set generated for the AMELI project (Alfons et al., 2011) and the simulation universe for the German census 2011 (see Münnich, Gabler, et al., 2012, Section 3.2 for details). An advantage of this strategy is that the true quantities we want to estimate, e.g. small area means, are known which makes it easy to assess their quality. This procedure will be called a fully design-based simulation study. An advantage of this kind of setup is that it enables a comparison of different estimation techniques in a realistic setting. This comes at the price that adequate modelling is much more difficult than in a quasi-design-based setting. In some situations with scarce register information validating a specific model may be impossible after all, e.g. the study of Burgard et al. (2014). Hence, to disentangle the impact of different sampling schemes from model misspecification may be a formidable task. Furthermore, from a scientific viewpoint, it may be very difficult to generalise the findings from a fully design-based simulation study. This is due to the fact that the data at hand can be very peculiar. Hence, a strategy suitable for one data set, may perform very badly on another data set, even if the same kind of target variable is considered.

A brief summary of simulation frameworks is given in Table A.1.

Table A.1: Taxonomy of simulation frameworks for small area estimation

Setup	Finite Population	Sample	Benchmark
purely model-based	none	directly drawn from model	expectation under model
finite population model-based	drawn from the model in every replication	drawn from the population	finite population quantity
quasi design-based	drawn from the model once and then held fixed	drawn from the population by means of a probability sample	finite population quantity
fully design-based	taken from a register	drawn from the population by means of a probability sample	finite population quantity

A.2 Quality measures in simulation studies

Once a simulation study has been conducted, the question how to evaluate the results arises. With respect to the quality of the point estimates, we are typically interested in two things. Firstly, the estimates should not systematically under- or overestimate the truth and secondly, the estimates should not exhibit a large variation. The former issue is related to the bias of the estimates, whereas the latter is about the precision. To assess whether biases in an area are present, the Monte-Carlo relative bias defined as

$$\mathrm{RBias}_{d,MC} := \frac{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_{r,d} - \theta_{r,d}\right)}{\frac{1}{R}\sum_{r=1}^{R}\theta_{r,d}} \tag{A.2.1}$$

can be used. In equation (A.2.1), $\widehat{\theta}_{r,d}$ denotes the estimated value for domain d in replication r, whereas $\theta_{r,d}$ refers to the true value in domain d and replication r. Equation (A.2.1) can be motivated as the Monte-Carlo approximation to the true relative bias, i.e.

$$\mathrm{RBias}_d = \frac{\mathrm{E}(\widehat{\theta}_d) - \theta_d}{\theta_d} \approx \frac{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_{r,d} - \theta_{r,d}\right)}{\frac{1}{R}\sum_{r=1}^{R}\theta_{r,d}}. \tag{A.2.2}$$

Despite the fact that (A.2.1) is an approximation to the true quantity (A.2.2), it is usually called as the relative bias. Note that (A.2.1) may take any value on the real line, while a relative bias of zero is desirable as this implies that on average the true value is realised. An issue with (A.2.1) arises, if the denominator in (A.2.1) is close to zero. In this case,

even tiny absolute biases, $R^{-1}\sum_{r=1}^{R}\left(\widehat{\theta}_{r,d}-\theta_{r,d}\right)$, can result in huge relative biases. Under these circumstances it may be better to report the absolute bias.

It should be further noted that in some cases the bias of an estimator is known to be zero. As an example, suppose $X_1,\ldots,X_n$ are i.i.d. realisations of a random variable with mean μ and a finite variance. It is a well-known fact that the sample mean is an unbiased estimator of μ when the expectation is with respect to the model. Hence, in this case, there is no need to consider a Monte-Carlo simulation, but for the sake of exposition suppose that a simulation study is conducted. Despite the unbiasedness of the estimator, the Monte-Carlo relative bias computed via (A.2.1) will, in general, not be zero. The reason for this behaviour is that Monte-Carlo approximation uses only a finite number of replications, while there is an infinite number of possible realisations of the model. In this case, there will be a Monte-Carlo bias but it should not be concluded that the estimator has some, albeit small, bias. An alternative measure to analyse the reliability of point estimates is the bias ratio, which is given by

$$\text{BiasRatio}_{d,MC} := \frac{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_{r,d}-\theta_{r,d}\right)}{\sqrt{\frac{1}{R-1}\left(\widehat{\theta}_{r,d}-\overline{\widehat{\theta}}_d\right)^2}}, \tag{A.2.3}$$

with $\overline{\widehat{\theta}}_d$ as the mean of $\widehat{\theta}_{r,d}$ over all simulations. Note that instead of the Monte-Carlo standard deviation, the mean of the estimated standard deviations could be used as well.

A second quantity which is usually reported is the Monte-Carlo relative root mean squared error defined as

$$\text{RRMSE}_{d,MC} := \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_{r,d}-\theta_{r,d}\right)^2}}{\frac{1}{R}\sum_{r=1}^{R}\theta_{r,d}}. \tag{A.2.4}$$

In the same spirit as (A.2.1), (A.2.4) can be motivated as the Monte-Carlo approximation to the true relative root mean squared error. The RRMSE can take values between 0 and ∞, where a value of 0 is clearly the most desirable. This measure is typically computed for point estimates. For unbiased estimators and a sufficiently large number of replications the Monte-Carlo RRMSE will be very close to the relative standard error, which is an important criterion for many national statistical institutes, as only results with a relative standard error lower than some threshold are published. Moreover, for point estimates the joint analysis of the relative bias and RRMSE yields valuable insights. In many cases, there will be a trade-off between bias and variance such that a model-based method may suffer from some bias, but can still be preferable on grounds of the RRMSE when compared to a design-based method (cf. the discussion in Lehtonen & Veijanen, 2009, p. 225). Note that the expressions (A.2.1) to (A.2.4) simplify for a design-based simulation study in which the true parameter θ_d does not depend on the simulation run.

Besides the quality of point estimates, we are also interested in the quality of precision estimates. One option in this regard is to check whether the variance or MSE estimates are unbiased. This can be evaluated in a simulation study using the formula for the relative bias given in (A.2.1). It should be noted, however, that while the computation of a true value for a point estimate is straightforward in many cases, deriving the true

variance or MSE can be much more complicated. As an alternative, we might exploit the fact that MSE of the point estimates serves as a Monte-Carlo approximation to the true MSE. This approach has some weaknesses as well, since the MSE of the point is a second moment, which does not converge as fast as a first moment. A common way of dealing with this problem is to compute MSE estimates for, say, $R = 1000$ replications and to compare against an estimate using the MSE of the point estimates which has been calculated using $R = 10000$ replications. However, substituting estimated quantities for unknown true values should only be carried out if no other option is possible.

Another approach to assess the quality of precision estimates is based on confidence intervals. Here, the basic notion is that the precision estimates should be such that the actual coverage rate in a Monte-Carlo simulation should be close to the nominal coverage rate. To formalise this idea, we consider the average confidence interval coverage rate (ACR) given by

$$\mathrm{ACR}_{1-\alpha,d} := \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}\left(\theta_{r,d} \in CI_{(\widehat{\theta}_{r,d},1-\alpha)}\right), \tag{A.2.5}$$

which yields the proportion of samples for which the confidence interval covers the true value (cf. Särndal et al., 1992, Remark 2.11.3). In addition to the coverage rates also the average confidence interval length may be of interest, as it entails information about our certainty. The shorter the confidence interval, the higher our certainty that the true value will be close to the point estimate. Since the endpoints of the confidence intervals depend on estimated variances (MSEs for model-based estimators), the coverage rate provides an implicit check on the accuracy of the variance / MSE estimates. An issue that frequently occurs when employing model-based techniques in a design-based environment is that the nominal coverage rate is not met. While this is clearly an important issue in the case of undercoverage, the consequences of overcoverage are may not be as negative. An implication of overcoverage is that the MSE estimates used are not efficient. This can be a serious problem if other estimation techniques yield shorter but reliable confidence intervals. In some cases, however, the too conservative intervals obtained by a model-based procedure may still be much shorter than any of the design-based or model-assisted alternatives. Thus, the joint analysis of coverage rates and average confidence interval lengths may yield valuable insights.

Note that the requirement of valid confidence intervals which meet the nominal coverage rate can be violated, even if the relative bias of the variance estimates is close to zero. This issue may occur if the distribution of the variance estimates is not smooth and unimodal, but has spikes. Thus, the relative bias would indicate that the variance estimation worked fine, when in fact it did not. This is due to the fact that the relative bias is an unconditional measure, whereas the coverage in single sample depends on the actual variance estimate for that sample. Hence, the ACR gives a better picture in this case. An example where this problem occurs was given in Section 6.3

The quality measures discussed so far are all area-specific. For the presentation of the results, it might be more convenient to publish a single summary statistic, which translates D area measures into one number. For the relative bias, the mean absolute relative bias (MARB) is frequently used. It is given by:

$$\text{MARB} := \frac{1}{D}\sum_{d=1}^{D} |\text{RBias}_{d,MC}|. \tag{A.2.6}$$

To summarise the results on the RRMSE, the average RRMSE (ARRMSE) is often reported. It can be computed via

$$\text{ARRMSE} := \frac{1}{D}\sum_{d=1}^{D} \text{RRMSE}_{d,MC}. \tag{A.2.7}$$

Appendix B

Additional material for Chapter 4

Table B.1: cAIC for augmented modelling under the Asparouhov size measure with $\alpha = 2$

	Augmented variable				
	none	w_{dj}	p_{dj}	p_{dj}^{-1}	$\log(p_{dj})$
$\log g(\mathbf{y}\|\widehat{\boldsymbol{\xi}}(\mathbf{y}), \widehat{\mathbf{v}}(\mathbf{y}))$	-773.55	-673.19	-647.43	-659.42	-650.66
K	78.96	89.25	87.33	87.16	87.38
cAIC	1705.02	1524.89	1469.54	1493.16	1476.07

Table B.2: cAIC for augmented modelling under the Asparouhov size measure with $\alpha = 3$

	Augmented variable				
	none	w_{dj}	p_{dj}	p_{dj}^{-1}	$\log(p_{dj})$
$\log g(\mathbf{y}\|\widehat{\boldsymbol{\xi}}(\mathbf{y}), \widehat{\mathbf{v}}(\mathbf{y}))$	-768.43	-724.80	-715.88	-719.21	-716.40
K	78.59	84.51	83.62	83.64	83.70
cAIC	1694.04	1618.62	1598.99	1605.70	1600.20

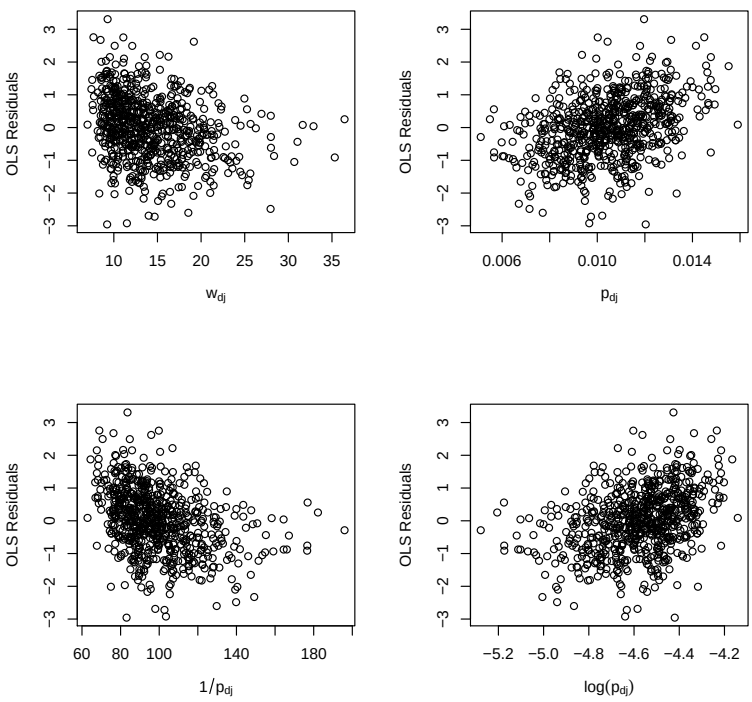

Figure B.1: OLS residuals plotted against various choices of the augmenting variable under the Asparouhov size measure with $\alpha = 2$

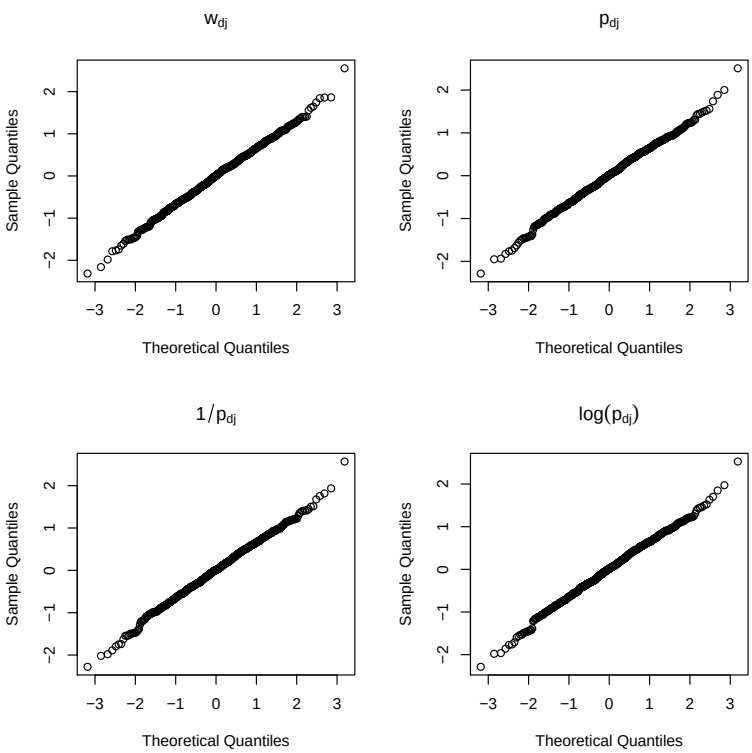

Figure B.2: QQ plots of the transformed residuals for various choices of the augmenting variable under the Asparouhov size measure with $\alpha = 2$

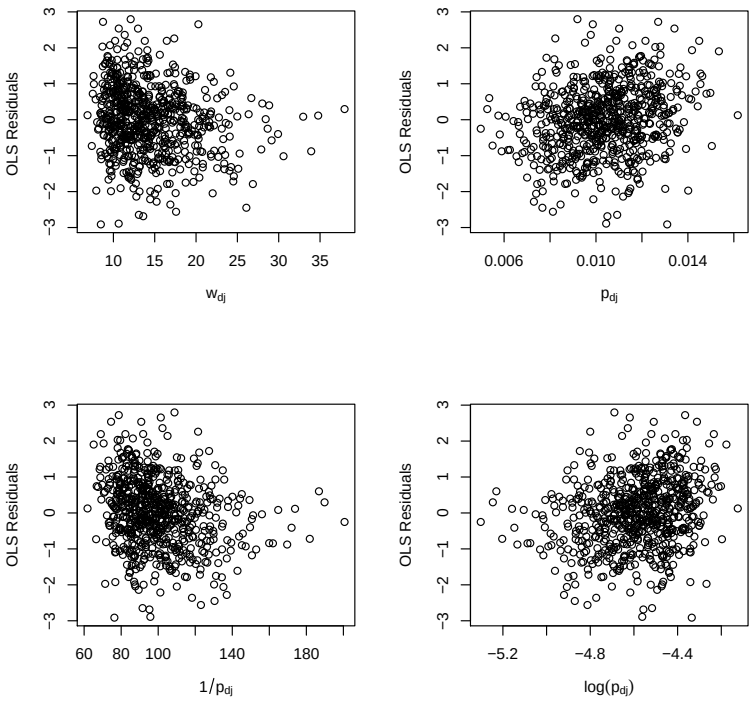

Figure B.3: OLS residuals plotted against various choices of the augmenting variable under the Asparouhov size measure with $\alpha = 3$

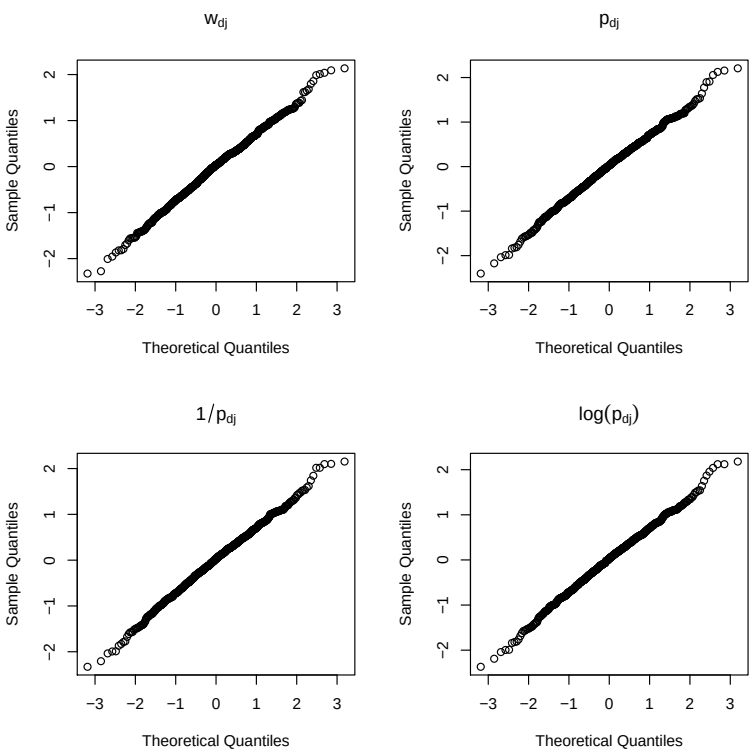

Figure B.4: QQ plots of the transformed residuals for various choices of the augmenting variable under the Asparouhov size measure with $\alpha = 3$

Appendix C

Additional material for Chapter 5

C.1 Antithetic clustering under a model with domain effects

To derive expression (5.2.29), we use (5.2.28) and plug these into equations (5.2.15). This yields:

$$\begin{aligned}
\mathrm{E}_M\left(\mathrm{SSB}_Y\right) &= \sum_{h=1}^{L} N_h \left(2\beta_0 \overline{V}_h + \beta_1^2 \overline{Z}_h^2 + 2\beta_1 \overline{Z}_h \overline{V}_h + \overline{V}_h^2 + \sigma^2/N_h\right) \\
&- N\left(\beta_0^2 + 2\beta_0\beta_1 \overline{Z} + 2\beta_0 \overline{V} + \beta_1^2 \overline{Z}^2 + 2\beta_1 \overline{ZV} + \overline{V}^2 + \sigma^2/N\right) \\
&= \beta_1^2 \underbrace{\left(\sum_{h=1}^{L} N_h \overline{Z}_h^2 - N\overline{Z}^2\right)}_{\mathrm{SSB}_Z} + 2\beta_1 \left(\sum_{h=1}^{L} N_h \overline{Z}_h \overline{V}_h - N\overline{ZV}\right) \\
&+ \underbrace{\sum_{h=1}^{L} N_h \overline{V}_h^2 - N\overline{V}^2}_{\mathrm{SSB}_Z} + (L-1)\,\sigma^2,
\end{aligned} \tag{C.1.1}$$

and

$$\begin{aligned}
\mathrm{E}_M\left(\mathrm{SSW}_Y\right) &= \sum_{k=1}^{N} \left(\beta_0^2 + 2\beta_0\beta_1 z_k + 2\beta_0 v_k + \beta_1^2 z_k^2 + 2\beta_1 z_k v_k + v_k^2 + \sigma^2\right) \\
&- \sum_{h=1}^{L} N_h \left(2\beta_0 \overline{V}_h + \beta_1^2 \overline{Z}_h^2 + 2\beta_1 \overline{Z}_h \overline{V}_h + \overline{V}_h^2 + \sigma^2/N_h\right) \\
&= \beta_1^2 \underbrace{\left(\sum_{k=1}^{N} z_k^2 - \sum_{h=1}^{L} N_h \overline{Z}_h^2\right)}_{\mathrm{SSW}_Z} + 2\beta_1 \left(\sum_{k=1}^{N} z_k v_k - \sum_{q=1}^{L} N_h \overline{Z}_h \overline{V}_h\right) \\
&+ \underbrace{\sum_{k=1}^{N} v_k^2 - \sum_{h=1}^{L} N_h \overline{V}_h^2}_{\mathrm{SSW}_Z} + L\sigma^2,
\end{aligned} \tag{C.1.2}$$

Assuming that the cross-product terms involving the clustering variable and the domain effects are negligible the approximations (5.2.29) follow.

C.2 Model-based simulation study under model misspecification

Table C.1: MARBs under model misspecification for planned domains in a model-based simulation study

		Direct	GREG			BHF		
Design	n_d		A	B	C	A	B	C
ATC	6	0.001	0.001	0.000	0.001	0.007	0.011	0.005
	14	0.001	0.000	0.000	0.000	0.004	0.007	0.002
	30	0.000	0.000	0.000	0.000	0.002	0.004	0.001
SRS	6	0.001	0.001	0.001	0.001	0.007	0.009	0.005
	14	0.001	0.001	0.001	0.000	0.004	0.005	0.002
	30	0.001	0.000	0.000	0.000	0.002	0.003	0.001
StrRS	6	0.002	0.001	0.001	0.001	0.007	0.015	0.004
	14	0.001	0.001	0.001	0.000	0.004	0.021	0.002
	30	0.001	0.000	0.000	0.000	0.002	0.005	0.001
UPS	6	0.001	0.001	0.003	0.001	0.007	0.067	0.004
	14	0.000	0.000	0.001	0.000	0.004	0.062	0.002
	30	0.000	0.000	0.002	0.000	0.002	0.060	0.001

Table C.2: ARRMSE under model misspecification for planned domains in a model-based study

		Direct	GREG			BHF		
Design	n_d		A	B	C	A	B	C
ATC	6	0.109	0.105	0.076	0.074	0.087	0.096	0.067
	14	0.069	0.069	0.049	0.048	0.063	0.066	0.046
	30	0.047	0.047	0.034	0.033	0.045	0.043	0.033
SRS	6	0.152	0.105	0.133	0.074	0.087	0.102	0.067
	14	0.100	0.069	0.088	0.049	0.063	0.076	0.046
	30	0.069	0.047	0.060	0.033	0.045	0.056	0.033
StrRS	6	0.125	0.104	0.100	0.074	0.086	0.098	0.067
	14	0.075	0.069	0.056	0.049	0.063	0.070	0.046
	30	0.050	0.047	0.037	0.033	0.045	0.044	0.033
UPS	6	0.113	0.107	0.094	0.076	0.087	0.124	0.067
	14	0.074	0.070	0.060	0.050	0.063	0.100	0.046
	30	0.050	0.048	0.046	0.034	0.045	0.083	0.033

Table C.3: ACR under model misspecification for planned domains in a model-based study

		Direct	GREG			BHF		
Design	n_d		A	B	C	A	B	C
ATC	6	0.951	0.951	0.950	0.950	0.948	0.929	0.954
	14	0.951	0.951	0.951	0.950	0.952	0.966	0.954
	30	0.950	0.950	0.950	0.949	0.953	0.988	0.953
SRS	6	0.950	0.949	0.950	0.951	0.947	0.940	0.953
	14	0.951	0.951	0.949	0.949	0.953	0.949	0.955
	30	0.949	0.948	0.950	0.949	0.952	0.953	0.954
StrRS	6	0.939	0.938	0.938	0.939	0.949	0.933	0.953
	14	0.942	0.943	0.941	0.941	0.954	0.966	0.955
	30	0.948	0.949	0.949	0.948	0.953	0.988	0.954
UPS	6	0.950	0.945	0.947	0.945	0.947	0.902	0.953
	14	0.951	0.946	0.949	0.946	0.952	0.890	0.954
	30	0.952	0.946	0.937	0.947	0.953	0.838	0.954

Table C.4: National estimates under misspecification in a model-based simulation study

		Direct			GREG		
Design	Model	RBias	RRMSE	ACR	RBias	RRMSE	ACR
ATC	A	0.0001	0.0119	0.9523	0.0002	0.0119	0.9472
	B				0.0000	0.0084	0.9501
	C				0.0000	0.0083	0.9502
UPS	A	−0.0001	0.0126	0.9499	0.0000	0.0122	0.9453
	B				−0.0011	0.0122	0.9288
	C				0.0000	0.0087	0.9458
StrRS	A	−0.0001	0.0126	0.9520	−0.0001	0.0117	0.9504
	B				0.0000	0.0096	0.9500
	C				0.0000	0.0085	0.9444
SRS	A	0.0000	0.0171	0.9508	−0.0002	0.0118	0.9477
	B				0.0001	0.0150	0.9503
	C				0.0001	0.0085	0.9488

Bibliography

Alfons, A., Filzmoser, P., Hulliger, B., Kolb, J.-P., Kraft, S., Münnich, R., & Templ, M. (2011). *Synthetic data generation of SILC data* (Research Project Report Nos. WP6 – D6.2). FP7-SSH-2007-217322 AMELI. Retrieved from http://ameli.surveystatistics.net

Alfons, A., Templ, M., & Filzmoser, P. (2010). An object-oriented framework for statistical simulation: The R package simFrame. *Journal of Statistical Software*, *37*(3), 1–36. Retrieved from http://www.jstatsoft.org/v37/i03/ doi: 10.18637/jss.v037.i03

Arora, V., & Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, *7*(4), 1053–1063.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics—Theory and Methods*, *35*(3), 439–460. doi: 10.1080/03610920500476598

Bankier, M. D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, *42*(3), 174-177. Retrieved from http://www.jstor.org/stable/2684995

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83 (401)*, 28-36. doi: 10.1080/01621459.1988.10478561

Berg, E., & Chandra, H. (2014). Small area prediction for a unit-level lognormal model. *Computational Statistics & Data Analysis*, *78*, 159–175. doi: 10.1016/j.csda.2014.03.007

Bernardini Papalia, R., Bruch, C., Enderle, T., Falorsi, S., Fasulo, A., Hernandez-Vazquez, E., ... Zimmermann, T. (2013). *Best practice recommendations on variance estimation and small area estimation in business surveys* (Tech. Rep.). Retrieved from http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable6.2.pdf

Booth, J. G., & Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, *93*(441), 262–272. doi: 10.1080/01621459.1998.10474107

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces.* John Wiley & Sons.

Brown, G., Chambers, R., Heady, P., & Heasman, D. (2001). Evaluation of small area estimation methods: An application to unemployment estimates from the UK LFS.

In *Proceedings of Statistics Canada Symposium.*

Burgard, J. P. (2013). *Evaluation of small area techniques for applications in official statistics* (Unpublished doctoral dissertation). University of Trier.

Burgard, J. P., Münnich, R., & Zimmermann, T. (2014). The impact of sampling designs on small area estimates for business data. *Journal of Official Statistics*, *30*(4), 749 - 771. doi: 10.2478/jos-2014-0046

Burgard, J. P., Münnich, R., & Zimmermann, T. (2015). Impact of sampling designs in small area estimation with applications to poverty measurement. In M. Pratesi (Ed.), *Analysis of poverty data by small area methods, forthcoming.* Wiley.

Cassel, C. M., Särndal, C. E., & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, *63*(3), 615–620.

Chambers, R. (2003). Introduction to part a. In R. L. Chambers & C. J. Skinner (Eds.), *Analysis of survey data* (pp. 13–27). John Wiley & Sons.

Chambers, R., Dorfman, A., & Sverchkov, M. (2003). Nonparametric regression with complex survey data. In R. L. Chambers & C. J. Skinner (Eds.), *Analysis of survey data* (pp. 151–174). John Wiley & Sons, Ltd.

Chambers, R., & Skinner, C. (2003a). *Analysis of survey data.* John Wiley & Sons.

Chambers, R., & Skinner, C. (2003b). Introduction. In R. Chambers & C. Skinner (Eds.), *Analysis of survey data* (pp. 1–10). John Wiley & Sons, Ltd. doi: 10.1002/0470867205.ch1

Chandra, H., & Chambers, R. (2011). Small area estimation under transformation to linearity. *Survey Methodology*, *37*, 39-51.

Chiodini, P. M., Martelli, B. M., Manzi, G., & Verrecchia, F. (2010). Between theoretical and applied approach: which compromise for unit allocation in business surveys? In *SIS conference.* Retrieved from https://www.researchgate.net/publication/228639697_Between_theoretical_and_applied_approach_which_compromise_for_unit_allocation_in_business_surveys

Choudhry, G. H., Rao, J. N. K., & Hidiroglou, M. A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, *38*(1), 23–29.

Costa, A., Satorra, A., & Ventura, E. (2004). Improving both domain and total area estimation by composition. *Statistics and Operations Research Transactions*, *28*(1), 69-86.

Crainiceanu, C. M., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(1), 165–185. doi: 10.1111/j.1467-9868.2004.00438.x

Das, K., Jiang, J., & Rao, J. N. K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, *32*(2), 818–840.

Datta, G. S., Hall, P., & Mandal, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, *106*(493), 362–374. doi: 10.1198/jasa.2011.tm10036

Datta, G. S., Kubokawa, T., Molina, I., & Rao, J. N. K. (2011). Estimation of mean squared error of model-based small area estimators. *Test*, *20*(2), 367–388. doi: 10.1007/s11749-010-0206-2

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.

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. doi: 10.1093/biomet/92.1.183

Eurostat. (2001). *The European Union labour force survey - methods and definitions - 2001.* Retrieved from http://epp.eurostat.ec.europa.eu/portal/page/portal/employment_unemployment_lfs/documents/EU_LFS_Methods_and_Definitions_2001.pdf (Accessed on 11/04/2014)

Falorsi, P. D., & Righi, P. (2008). A balanced sampling approach for multi-way stratification designs for small area estimation. *Survey Methodology*, *34*(2), 223-234.

Fay, R. E., & Herriot, R. A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74 (366)*, 269-277. doi: 10.1080/01621459.1979.10482505

Foster, J., Greer, J., & Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, 761–766. doi: 10.2307/1913475

Fuller, W. A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological, and Environmental Statistics*, 331–345.

Fuller, W. A. (2009). *Sampling statistics.* John Wiley & Sons.

Gabler, S., Ganninger, M., & Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika*, *75*(2), 151–161. doi: 10.1007/s00184-010-0319-3

Gabler, S., Häder, S., & Lynn, P. (2006). Design effects for multiple design samples. *Survey Methodology*, 115–120.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*, 153-164. doi: 10.1214/088342306000000691

Ghosh, M., Natarajan, K., Stroud, T., & Carlin, B. P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, *93*(441), 273–282. doi: 10.1080/01621459.1998.10474108

González-Manteiga, W., Lombardía, M., Molina, I., Morales, D., & Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, *78*(5), 443–462. doi: 10.1080/00949650601141811

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational statistics & data analysis*, *51*(5), 2720–2733. doi: 10.1016/j.csda.2006.01.012

Graf, M., Wenger, A., & Nedyalkova, D. (2011). *Description and quality of the user data base* (Research Project Report Nos. WP5 – D5.1). FP7-SSH-2007-217322 AMELI.

Retrieved from http://ameli.surveystatistics.net

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional aic in linear mixed models. *Biometrika*, 773-789. doi: 10.1093/biomet/asq042

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849-879. doi: 10.1093/poq/nfq065

Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society*, *68 (2)*, 221-238. doi: 10.1111/j.1467-9868 .2006.00541.x

Han, B. (2013). Conditional Akaike information criterion in the Fay–Herriot model. *Statistical Methodology*, *11*, 53–67. doi: 10.1016/j.stamet.2012.09.002

Hardy, G. H., Littlewood, J. E., & Pólya, G. (1952). *Inequalities*. Cambridge university press.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 153–161. doi: 10.2307/1912352

Hidiroglou, M. A., & Lavallee, P. (2009). Sampling and estimation in business surveys. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics* (Vol. 29 A, p. 441-470). New York: Elsevier. doi: 10.1016/S0169-7161(08)00017-5

Hidiroglou, M. A., & Patak, Z. (2004). Domain estimation using linear regression. *Survey Methodology*, *30*, 67–78.

Huang, R., & Hidiroglou, M. (2003). Design consistent estimators for a mixed linear model on survey data. *Proceedings of the Survey Research Methods Section, American Statistical Association (2003)*, 1897–1904.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer.

Jiang, J., & Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, *53*(2), 217-243. doi: 10.1023/A:1012410420337

Jiang, J., & Lahiri, P. (2006a). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*. doi: 10.1198/016214505000000790

Jiang, J., & Lahiri, P. (2006b). Mixed model prediction and small area estimation. *Test*, *15 (1)*, 1-96. doi: 10.1007/BF02595419

Jiang, J., Lahiri, P., Wan, S.-M., et al. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics*, *30*(6), 1782–1810. doi: doi:10.1214/aos/1043351257

Jiang, J., Lahiri, P., Wan, S.-M., & Wu, C.-H. (2001). Jackknifing in the Fay-Herriot model with an example. In *Proceedings of the seminar on funding opportunity in survey research.*

Kackar, R. N., & Harville, D. A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in statistics-theory and methods*, *10*(13), 1249–1261. doi: 10.1080/03610928108828108

Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effect in mixed linear models. *Journal of the American Statistical Association*, *79*(388), 853-862. doi: 10.1080/01621459.1984.10477102

Karlberg, F. (2000). Population total prediction under a lognormal superpopulation model. *Metron*, *58*(3/4), 53–80.

Kennel, T. L., & Valliant, R. (2010). Logistic generalized regression (LGREG) estimator in cluster samples. In *Survey research methods* (pp. 4756–4770).

Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. John Wiley & Sons.

Korn, E. L., & Graubard, B. I. (1999). *Analysis of health surveys*. John Wiley & Sons.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, *18*, 199-210. doi: doi:10.1214/ss/1063994975

Lahiri, P., & Rao, J. N. K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, *90*(430), 758-766. doi: 10.1080/01621459.1995.10476570

Lahiri, S. N., Maiti, T., Katzoff, M., & Parsons, V. (2007). Resampling-based empirical prediction: an application to small area estimation. *Biometrika*, *94*(2), 469-485. doi: 10.1093/biomet/asm035

Lehtonen, R., & Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys*. John Wiley & Sons.

Lehtonen, R., Särndal, C. E., & Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, *29*(1), 33-44.

Lehtonen, R., Särndal, C. E., & Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, *7*(3), 649-673.

Lehtonen, R., & Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, *24*(1), 51-55.

Lehtonen, R., & Veijanen, A. (2009). Design-based methods of estimation for domains and small areas. In D. Pfeffermann & C. Rao (Eds.), *Handbook of statistics* (Vol. 29 B, p. 219 -249). New York: Elsevier.

Lehtonen, R., Veijanen, A., Myrskylä, M., & Valaste, M. (2011). *Small area estimation of indicators on poverty and social exclusion* (Tech. Rep.). AMELI deliverable D2.2. Retrieved from https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Delivrables/AMELI-WP2-D2.2-20110402.pdf

Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of multivariate analysis*, *101*(4), 882–892. doi: https://doi.org/10.1016/j.jmva.2009.10.009

Little, R. J. (2003). The Bayesian approach to sample survey inference. In R. L. Chambers & C. J. Skinner (Eds.), *Analysis of survey data* (pp. 49–57). Wiley Online Library.

Little, R. J. (2012). Calibrated Bayes, an alternative inferential paradigm for official

statistics. *Journal of Official Statistics*, *28*(3), 309–334.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons.

Liu, B. (2009). *Hierarchical Bayes estimation and empirical best prediction of small-area proportions* (Doctoral dissertation, University of Maryland). Retrieved from http://drum.lib.umd.edu/bitstream/1903/9149/1/Liu_umd_0117E_10245.pdf

Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.

Lohr, S., & Rao, J. N. K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika*, 457-468. doi: 10.1093/biomet/asp003

Longford, N. T. (2001). Simulation-based diagnostics in random-coefficient models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *164*(2), 259–273. doi: 10.1111/1467-985X.00201

Longford, N. T. (2006). Sample size calculation for small area estimation. *Survey Methodology*, *32*(1), 87-96.

López-Vizcaíno, E., Lombardía, M. J., & Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, *13*(2), 153–178. doi: 10.1177/1471082X13478873

Maiti, T. (2004). Applying jackknife method of mean squared prediction error estimation in SAIPE. *Statistics in Transition*, *6*(5), 685–695.

Maiti, T., & Slud, E. V. (2002). *Comparison of small area models in SAIPE* (Tech. Rep.). Technical Report Census Bureau.

Marker, D. A. (1995). *Small area estimation: A Bayesian perspective*. University of Michigan.

Marker, D. A. (2001). Producing small area estimates from national surveys: Methods for minimizing use of indirect estimators. *Survey Methodology*, *27*(2), 183-188.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall.

Meinfelder, F. (2014). Multiple imputation: an attempt to retell the evolutionary process. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *8*, 249–267. doi: 10.1007/s11943-014-0151-8

Molefe, W. B. (2011). *Sample design for small area estimation* (Doctoral dissertation, University of Wollongong). Retrieved from http://ro.uow.edu.au/cgi/viewcontent.cgi?article=4497&context=theses

Molina, I., & Marhuenda, Y. (2013). sae: Small area estimation [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=sae (R package version 1.0-2)

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, *38*(3), 369–385. doi: 10.1002/cjs.10051

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. doi: 0.1111/j.1467-985X .2007.00493.x

Molina, I., Salvati, N., & Pratesi, M. (2009). Bootstrap for estimating the mse of the spatial eblup. *Computational Statistics*, *24*(3), 441–458. doi: 10.1007/s00180-008 -0138-4

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, *28*(2), 135–167. doi: doi:10.1214/12-STS410

Münnich, R., & Burgard, J. P. (2012). On the influence of sampling design on small area estimates. *Journal of the Indian Society of Agricultural Statistics*, *66*(1), 145-156.

Münnich, R., Gabler, S., Ganninger, M., Burgard, J. P., & Kolb, J.-P. (2012). *Stichprobenoptimierung und Schätzung im Zensus 2011.* Statistisches Bundesamt.

Münnich, R., Sachs, E., & Wagner, M. (2012). Calibration benchmarking for small area estimates: an application to the German census 2011. In *Fields institute symposium on the analysis of survey data and small area estimation in honour of the 75th birthday of JNK Rao, Ottawa, Canada.*

Münnich, R., Sachs, E., & Wagner, M. (2012). Numerical solution of optimal allocation problems in stratified sampling under box constraints. *Advances in Statistical Analysis*, *96*(3), 435-450. doi: 10.1007/s10182-011-0176-z

Myrskylä, M. (2007). *Generalised regression estimation for domain class frequencies* (Doctoral dissertation, University of Helsinki). Retrieved from http://helda.helsinki .fi/bitstream/handle/10138/23380/generali.pdf?sequence=1

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal Of The Royal Statistical Society*, *97*, 558-625. doi: 10.2307/2342192

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., & Breidt, F. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 265–286. doi: 10.1111/j.1467-9868.2007.00635.x

Opsomer, J. D., Francisco-Fernández, M., & Li, X. (2012). Model-based non-parametric variance estimation for systematic sampling. *Scandinavian Journal of Statistics*, *39*(3), 528–542. doi: 10.1111/j.1467-9469.2011.00773.x

Pereira, L. N., Mendes, J. M., & Coelho, P. S. (2013). Model-based estimation of unemployment rates in small areas of portugal. *Communications in Statistics-Theory and Methods*, *42*(7), 1325–1342. doi: 10.1080/03610926.2012.749989

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, *61*(2), 317-337. doi: 10.2307/1403631

Pfeffermann, D. (2002). Small area estimation – new developments and directions. *International Statistical Review*, *70*(1), 125-143. doi: 10.1111/j.1751-5823.2002.tb00352 .x

Pfeffermann, D. (2006). Comment on: Mixed model prediction and small area estimation. *Test*, *1*(1), 65-72.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it. *Survey Methodology*, *37*(2), 115–136.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, *28*(1), 40–68. doi: 10.1214/12-STS395

Pfeffermann, D. (2014). *Small area estimation: Model selection and checking.* The international conference on Small Area Estimation in Poznan.

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. doi: 10.1093/biomet/ass010

Pfeffermann, D., & Glickman, H. (2004). Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap. In *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Pfeffermann, D., & Rao, C. R. (2009a). *Handbook of Statistics: Sample Surveys: Design, Methods and Applications* (Vol. 29A). Elsevier.

Pfeffermann, D., & Rao, C. R. (2009b). *Handbook of Statistics: Sample Surveys: Inference and Analysis* (Vol. 29B). Elsevier.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilites in multilevel models. *Journal Of The Royal Statistical Society Series B*, *60*(1), 23-40. doi: 10.1111/1467-9868.00106

Pfeffermann, D., & Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, *102*(480), 1427-1439. doi: 10.1198/016214507000001094

Pfeffermann, D., & Sverchkov, M. (2009). Inference under informative sampling. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics* (Vol. 29 B, p. 455-487). New York: Elsevier.

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. doi: 10.1080/01621459.1990.10475320

Prasad, N. G. N., & Rao, J. N. K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology*, *25*(1), 67-72.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(4), 805–827. doi: 10.1111/j.1467-985X.2006.00426.x

Rao, J. N. K. (2003). *Small area estimation.* New York: John Wiley and Sons.

Rao, J. N. K. (2007). Jackknife and bootstrap methods for small area estimation. In *Proceedings of the survey research methods section, American Statistical Association* (p. 2925 - 2929).

Rao, J. N. K., Hidiroglou, M., Yung, W., & Kovacevic, M. S. (2010). Role of weights in descriptive and analytical inferences from survey data: An overview. *Journal of the Indian Society of Agricultural Statistics*, *64*, 129 - 135.

Rao, J. N. K., Verret, F., & Hidiroglou, M. A. (2013). A weighted composite likelihood

approach to inference for two-level models from survey data. *Survey Methodology*, *39*(2), 263 - 282.

Rizzo, M. L. (2007). *Statistical computing with R.* CRC Press.

Saei, A., & Taylor, A. (2012). Labour force status estimates under a bivariate random components model. *Journal of the Indian Society of Agricultural Statistics*.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, *33*(2), 99–119.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling.* Springer.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components.* John Wiley & Sons.

Sinha, S. K., & Rao, J. N. K. (2009). Robust small area estimation. *Canadian Journal of Statistics*, *37*(3), 381–399. doi: 10.1002/cjs.10029

Skinner, C. (1994). Sample models and weights. In *Proceedings of the Section on Survey Research Methods* (pp. 133–142).

Slud, E. V., & Maiti, T. (2006). Mean-squared error estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 239–257. doi: 10.1111/j.1467-9868.2006.00542.x

Tillé, Y. (2006). *Sampling algorithms.* Springer.

Torabi, M., & Rao, J. N. K. (2008). Small area estimation under a two-level model. *Survey Methodology*, *34*(1), 11.

Torabi, M., & Rao, J. N. K. (2010). Mean squared error estimators of small area means using survey weights. *Canadian Journal of Statistics*, *38*, 598-608. doi: 10.1002/cjs.10078

Tschuprow, A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, *2*, 461-493.

Tse, T., & Esposito, M. (2013). *Youth unemployment could wreck europe's economic recovery.* Retrieved from http://www.theguardian.com/commentisfree/2013/nov/14/youth-unemployment-wreck-europe-economic-recovery (Downloaded on June 29th, 2015)

Vaida, F., & Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, *92*(2), 351–370. doi: 10.1093/biomet/92.2.351

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling and inference: a prediction approach.* John Wiley.

Verbeek, M. (2008). *A guide to modern econometrics.* John Wiley & Sons.

Verret, F., Hidiroglou, M. A., & Rao, J. N. K. (2015). Model-based small area estimation under informative sampling. *Survey Methodology*, *41*(2), 333–347.

Wang, J., Fuller, W. A., & Qu, Y. (2008). Small area estimation under a restriction. *Survey Methodology*, *34*(1), 29-36.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated

estimate. *Journal of the American Statistical Association*, *66*(334), 411–414.

You, Y., & Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, *32*(1), 97-103.

You, Y., & Rao, J. N. K. (2002a). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, *30*, 431-439. doi: 10.2307/3316146

You, Y., & Rao, J. N. K. (2002b). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, *30*(1), 3–15. doi: 10.2307/3315862

You, Y., Rao, J. N. K., & Hidiroglou, M. (2013). On the performance of self benchmarked small area estimators under the Fay-Herriot area level model. *Survey Methodology*, *39*, 217–229.

Zhang, L.-C. (2007). Finite population small area interval estimation. *Journal of Official Statistics*, *23*(2), 223-237.

Zhang, L.-C. (2009). Estimates for small area compositions subjected to informative missing data. *Survey Methodology*, *35*(2), 191 - 201.

Zhang, L.-C., & Chambers, R. L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(2), 479–496. doi: 10.1111/j.1369-7412.2004.05266.x

Zimmermann, T. (2017). Variance reduction using a non-informative sampling design. *Statistical Journal of the IAOS*, Preprint. doi: 10.3233/SJI-170358