Finite Mixture Models for Small Area Estimation in Cases of Unobserved Heterogeneity

Submitted in partial fulfillment of the requirements for the degree

Dr. rer. pol.

to the

Department IV
University of Trier

Charlotte Articus
Clarenbachstraße 190
50931 Köln

Supervisors:
Prof. Dr. Ralf Münnich (University of Trier)
Prof. Daniela Cocchi, PhD (University of Bologna)

April 2018

ACKNOWLEDGEMENTS

First of all, I want to thank my first supervisor Prof. Dr. Ralf Münnich – not only for his valuable scientific guidance but also for his confidence in me and his creative and very generous way of finding solutions that allowed to bring my family and my career in line. This work would not have been possible without his ongoing and sympathetic support in this regard.

I also want to thank Prof. Daniela Cocchi for agreeing to supervise my thesis and for a warm-hearted defense with a scientifically interesting discussion. I really enjoyed this interchange.

My work was supported by the Nikolaus Koch Stiftung through the REMIKIS project and by the European Union's Seventh Framework Programme for Research, Technological Development and Demonstration under Grant No. 312691 (InGRID). I am grateful for this support. Further, I thank the Federal Statistical Office of Germany for kindly providing the survey data on rental prices used in this study.

I thank my colleagues at the Economic and Social Statistics Department for a great and cooperative working atmosphere. I am particularly grateful to my friend Lisa Borsi, with whom I shared a small rooftop office for some months. She was always willing to discuss my questions and repeatedly surprised me by her permanent readiness and ability to immediately think through my questions and provide valuable support. I am also grateful to Dr. Pablo Burgard who had the initial idea of introducing mixtures into SAE. He therewith laid the foundations for this research and this thesis would not have been written in this form without him. I also thank him for many lively discussions that really helped to clarify unsolved issues and to carry ideas further.

Finally, I would like to thank my family and friends for the strong background, the ongoing support and welcome diversion. I particularly thank my husband Malte who was an ever-supportive partner in this project and who, as an involved father, was always willing to take responsibility for our children to facilitate my work.

Abstract

A basic assumption of standard small area models is that the statistic of interest can be modelled through a linear mixed model with common model parameters for all areas in the study. The model can then be used to stabilize estimation. In some applications, however, there may be different subgroups of areas, with specific relationships between the response variable and auxiliary information. In this case, using a distinct model for each subgroup would be more appropriate than employing one model for all observations. If no suitable natural clustering variable exists, finite mixture regression models may represent a solution that 'lets the data decide' how to partition areas into subgroups. In this framework, a set of two or more different models is specified, and the estimation of subgroup-specific model parameters is performed simultaneously to estimating subgroup identity, or the probability of subgroup identity, for each area. Finite mixture models thus offer a flexible approach to accounting for unobserved heterogeneity.

Therefore, in this thesis, finite mixtures of small area models are proposed to account for the existence of latent subgroups of areas in small area estimation. More specifically, it is assumed that the statistic of interest is appropriately modelled by a mixture of K linear mixed models. Both mixtures of standard unit-level and standard area-level models are considered as special cases. The estimation of mixing proportions, area-specific probabilities of subgroup identity and the K sets of model parameters via the EM algorithm for mixtures of mixed models is described. Eventually, a finite mixture small area estimator is formulated as a weighted mean of predictions from model 1 to K, with weights given by the area-specific probabilities of subgroup identity.

Finite mixture models have been extended to include additional covariates to model the mixture weights. This is particularly useful if the aim is not only to control for heterogeneity as a nuisance in the data but also to identify and characterize the subgroups in a meaningful way. If suitable covariates are available, the submodel also supports the assignment to subgroups. Moreover, it can be used to classify new observations on the basis of the covariates alone. Therefore, a corresponding extension for the finite mixture of small area models is also considered. In addition to the advantages listed above, in a small area context, the improved assignment to subgroups also enhances the accuracy of the estimation of the statistic of interest. Furthermore, the option of assigning new observations

to subgroups based on the estimated submodel and the covariates only can be employed to predict the statistic of interest for unsampled areas in a heterogeneous population.

The approach suggested in this work is inspired by an attempt to estimate regional rental prices in Germany on the basis of the German *Mikrozensus*. This important household survey, which is conducted by the Federal Statistical Institute, periodically contains a special section on housing. It therewith provides otherwise scarce nationwide information on rental prices. Due to precision requirements, however, results on average rents are only published at the level of the German Länder. As rental prices vary significantly between regions, this level of aggregation is far too high for many purposes. Therefore, in a first step, a standard area-level model was estimated in order to use the information provided by the survey efficiently. Regional indicators such as the population growth rate, the prevalence of rented housing and the price of building land where used as covariates. Overall, reliable results on average rental prices on district level were obtained. The application did, however, raise doubts as to whether it is appropriate to assume one model for all areas. Factors that drive rental markets very likely vary between different types of areas, such as rural and urban districts. These concerns motivated the proposal of a mixture-based approach to small area estimation. Furthermore, the need to gain a deeper understanding of the segmentation, inspired the incorporation of a submodel for the mixture weights into the framework.

While the proposal is motivated by this specific application, it could also be an appropriate method in any application in which small area estimates for heterogeneous subentities are of interest. Furthermore, the suggested estimator could also be interpreted as a flexible approach when the distribution of the statistic of interest is unknown and the usual normality assumption of the basic small area models seems inappropriate.

The proposed method is evaluated in model-based simulation studies. It is then applied to the problem of estimating rental prices at the district level in Germany.

Zusammenfassung

Eine zentrale Annahme small-area-statistischer Standardmodelle ist, dass die interessierende Variable durch ein Lineares Gemischtes Modell modelliert werden kann. Die Modellparameter sind dabei für alle *areas* gleich. In einigen Anwendungen scheint es jedoch plausibler, dass es verschiedene Gruppen von *areas* mit jeweils spezifischem Zusammenhang zwischen interessierender Variable und Kovariablen gibt. In einem solchen Fall unbeobachteter Heterogenität wäre ein eigenes Modell für jede Gruppe angemessener als ein gemeinsames Modell für alle *areas*. Wenn die Gruppen jedoch unbeobachtet sind und es keine geeignete natürliche Clustering-Variable gibt, kann die Schätzung eines Finite Mixture Models eine geeignete Methode sein, um die Regionen auf Grundlage der verfügbaren Daten in Gruppen zu unterteilen. Dazu werden zwei oder mehr verschiedene Modelle spezifiziert. Die Schätzung der gruppenspezifischen Modellparameter sowie der Gruppenzugehörigkeit der einzelnen *areas* erfolgt dann simultan. Finite Mixture Models sind damit ein flexibler methodischer Ansatz um unbeobachtete Heterogenität zu berücksichtigen.

In dieser Arbeit wird daher eine Finite Mixtures von Small Area Modellen vorgeschlagen, um latente Gruppen in small-area-statistischen Anwendungen zu berücksichtigen. Konkret wird angenommen, dass die interessierende Variable durch eine Mischung von K gemischten Modellen modelliert werden kann. Sowohl das Standard Unit-Level als auch das Standard Area-Level Modell werden als Spezialfälle betrachtet. Die Schätzung der Mischungsgewichte, der areaspezifischen Wahrscheinlichkeiten für Gruppenzugehörigkeit und der K Vektoren der Modellparameter erfolgt über den EM-Algorithmus. Schließlich, wird ein Finite Mixture Small Area Schätzer als gewichtetes Mittel der Prädiktionen aus den Modellen 1 bis K formuliert. Die Gewichte sind dabei die *area*-spezifischen Wahrscheinlichkeiten für Komponentenzugehörigkeit.

Finite Mixture Models können um ein Modell für die Mischungsgewichte erweitert werden. Das ist vor allem dann sinnvoll, wenn es nicht nur um eine Kontrolle von unbeobachteter Heterogenität als Störung in den Daten geht, sondern die Gruppen auch identifiziert und inhaltlich interpretiert werden sollen. Wenn geeignete Kovariablen verfügbar sind, unterstützt das Untermodell auch die Zuordnung zu den Komponenten. Außerdem kann es verwendet werden, um neue Beobachtungen auf Grundlage der Kovariablen zu klassifizieren. Daher wird eine entsprechende Erweiterung auch für Finite Mixtures von Small Area Modellen betrachtet. Zusätzlich zu den bereits genannten Vorzügen, wirkt sich die verbesserte

Schätzung der Komponentenzugehörigkeiten im Small Area Kontext auch positiv auf die Präzision der Prädiktion der interessierenden Variable aus. Außerdem kann die Zuordnung neuer Beobachtungen auf Basis der verfügbaren Kovariablen genutzt werden, um die interessierende Variable von *areas* mit einem Stichprobenumfang von null in einer heterogenen Population zu prädizieren.

Der vorgeschlagene Ansatz ist durch eine Anwendung small-area-statistischer Verfahren für die Schätzung von Mietpreisen auf Grundlage des Mikrozensus inspiriert. Dieser wichtige Haushaltssurvey des Statistischen Bundesamtes enthält alle vier Jahre eine Zusatzerhebung zur Wohnsituation und stellt damit ansonsten nicht verfügbare, flächendeckende Informationen zu Bestandmieten in Deutschland zur Verfügung. Aufgrund von Präzisionsanforderungen werden Auswertungen über mittlere Mieten jedoch nur auf Ebene der Bundesländer zur Verfügung gestellt. Mietpreise schwanken allerdings deutlich zwischen den Regionen, so dass dieses Aggregationslevel für viele Verwendungszwecke deutlich zu hoch ist. Daher wurde in einem ersten Schritt ein Standard Area-Level Modell geschätzt, um die im Mikrozensus verfügbaren Informationen effizient zu nutzen. Als Kovariablen wurden Regionalindikatoren wie die Bevölkerungsentwicklung, die Bedeutung des Mietmarktes und Baulandpreise verwendet. Insgesamt konnten verlässliche Ergebnisse über durchschnittliche Mieten auf Kreisebene gewonnen werden. Die Anwendung ließ aber zweifeln, ob die Annahme eines gemeinsamen Modells für alle *areas* angemessen ist: Es scheint plausibler, dass Determinanten der Mietpreisbildung zwischen verschiedenen Typen von *areas*, zum Beispiel zwischen urbanen und ländlichen Kreisen, divergieren. Diese Bedenken motivierten einen mischungsbasierten Ansatz für Small Area Statistik. Das Interesse an einem tieferen Verständnis der Segmentierung inspirierte die Erweiterung um ein Untermodell für die Mischungsgewichte.

Obwohl der vorgeschlagene Ansatz durch diese spezifische Anwendung motiviert worden ist, können Finite Mixtures von Small Area Modellen selbstverständlich auch in anderen Anwendungen, in denen Small Area Schätzungen für eine heterogene Population interessieren, eine geeignete Methode darstellen. Der vorgeschlagene Schätzer kann außerdem als ein flexibler Ansatz für Anwendungen interpretiert werden, in denen die Verteilung der interessierenden Variable unbekannt ist und die üblichen Normalverteilungsannahmen von Standardmodellen der Small Area Statistik nicht angemessen erscheinen.

Der vorgeschlagene Ansatz wird in modellbasierten Simulationsstudien evaluiert. Anschließend wird das Verfahren auf die Schätzung von regionalen Mietpreisen in Deutschland angewendet.

Contents

List of Figures

List of Tables

List of Notations and Notational Remarks

$\#(\boldsymbol{a})$	Cardinality of $\boldsymbol{a}$
$\lvert\boldsymbol{A}\rvert$	Determinant of matrix $\boldsymbol{A}$
$\boldsymbol{A}^T$, $\boldsymbol{a}^T$	Transpose of a matrix $\boldsymbol{A}$ or a vector $\boldsymbol{a}$, respectively
$\text{diag}(a_1, \ldots, a_n)$	Diagonal matrix with diagonal elements $a_1, \ldots, a_n$
$\text{diag}(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_n)$	Block-diagonal matrix with square matrices $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_n$ in the main diagonal.
$\boldsymbol{I}_n$	$n \times n$ identity matrix
$\boldsymbol{\mathcal{I}}$	Fisher information matrix
$L(.)$	Likelihood function
$l(.)$	Log-likelihood
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\text{rank}(\boldsymbol{A})$	Rank of matrix $\boldsymbol{A}$
$\text{tr}(\boldsymbol{A})$	Trace of matrix $\boldsymbol{A}$
$\boldsymbol{1}_n$	n-vector of 1s

Generally, all vectors and matrices are printed in boldface. If not explicitly stated otherwise, vectors are represented by lowercase letters and matrices by capital letters.

List of Abbreviations

ARRMSE	Average Relative Root Mean Square Error
BHF	Battese Harter Fuller
BIC	Bayesian Information Criterion
BICadj	Sample-Size Adjusted Bayesian Information Criterion
BLUE	Best Linear Unbiased Estimator
BLUP	Best Linear Unbiased Predictor
BP	Best Predictor
BLP	Best Linear Predictor
CV	Coefficient of Variation
EBLUP	Empirical Best Linear Unbiased Predictor
EM	Expectation-Maximization
FH	Fay Herriot
FMM	Finite Mixture Model
GLMM	General Linear Mixed Model
GLS	Generalized Least Squares
ICL	Integrated Classification Likelihood
LMM	Linear Mixed Model
MARB	Mean Absolute Relative Bias
ML	Maximum Likelihood
MC	Monte Carlo
MSE	Mean Square Error
RBIAS	Relative Bias
REML	Restricted Maximum Likelihood
RMSE	Root Mean Square Error
RRMSE	Relative Root Mean Square Error
SAE	Small Area Estimation

Chapter 1

Introduction

When analysing survey data in order to gain insights into social or economic phenomena, the aim may not only be to make statistical inferences about the entire target population but also to obtain reliable information regarding certain subentities. Such subentities may be, for example, smaller areas in a region under study or demographic subgroups in a population. However, researchers are frequently confronted with the problem that the spatial or thematic disaggregation of the available sample results in small subsamples for the subentities of interest, which causes a lack of accuracy when using conventional direct estimators. Model-based Small Area Estimation (SAE) techniques, that are designed to produce reliable information even for small subsamples, may represent a solution. These methods are aimed at improving the efficiency of estimation in the case of small subsamples by means of an explicit statistical model. The intuition is that part of the variation in these statistics can be explained by a relationship between the variable of interest and a certain set of covariates that is valid for all areas under consideration. Thus, it can be estimated using the data points from all subsamples. The specified relationship can then be employed to stabilize the prediction of the variable of interest.

In many applications, however, it is plausible to assume that the relationship between the variable of interest and given auxiliary information will differ by area. It then seems reasonable to consider different subgroups of areas, and the estimation of subgroup-specific models might be more appropriate than using one model for all areas. There might be a natural clustering variable, which could be used to segment areas into two or more subgroups. If this is not the case, the definition of subgroups becomes a crucial task in the estimation process and appropriate techniques to partition the areas have to be applied. Finite mixture regression models seem to represent a natural solution to this problem. In this framework, a set of

two or more different models is specified, and the estimation of model parameters is performed simultaneously to estimating subgroup identity, or the probability of subgroup identity, for each area. Mixture models thus offer a flexible, integrated approach to accounting for latent subgroups in the population.

The objective of this thesis is to propose mixture-based small area estimators to account for the existence of unobserved subgroups of areas. More specifically, it is assumed that the observed values of a target statistic are appropriately modelled by a finite mixture of K small area models. Subgroup-specific model parameters and area-specific probabilities of subgroup membership are estimated simultaneously using the Expectation-Maximization (EM) algorithm. An estimator of the target statistic is then obtained as a weighted average of the predictions from the K component models. Weights are given by the area-specific probability of subgroup-membership. In a second step, the model is extended to include a (concomitant variable) submodel for the mixture weights in order to support clustering and to gain further insights into the clustered structure. Based on an investigation into relevant theory from the fields of both model-based SAE and finite mixture models (FMM), details on relevant model specifications, parameter estimation, and prediction are presented. More specifically, both mixture of unit-level and of area-level models, and respective estimators, are proposed. Furthermore, suitable approaches for determining the number of components are discussed.

The primary objective of the approaches suggested in this work is to improve estimation performance in cases of a clustered population by providing a model that fits the data structure. In addition, the model-based probabilistic clustering of areas, which is obtained as a by-product of the estimation process, may provide valuable insights into underlying patterns. It therefore furthers the understanding of the data at hand.

The proposal is inspired by and employed to the problem of estimating regional rental prices in Germany. Measuring rental prices is of high practical relevance, as they make up an important share of private households' living expenses and, as such, constitute a crucial determinant of consumer price indices. Moreover, rental prices provide valuable information concerning the situation of the housing market and thus indicate demand for political action and regional development planning. As regional rental markets develop in a highly heterogeneous manner, there is an interest in measuring prices at the local level. However, a comprehensive regional differentiation other than that at the level of the German *Länder* is usually not provided, as the ability to provide estimates at a higher level of disaggregation is restricted by rather small sample sizes for regional subentities. An exception are the quoted rents at district level, which are calculated by the BBSR (Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR), 2012). Therefore,

estimating regional rental prices constitutes a natural application for SAE.

The analysis in this thesis is based on data provided by the German *Mikrozensus*, an important household survey conducted by the Federal Statistical Institute, which contains a special section on housing every fourth year. It thus provides otherwise scarce nationwide information on rental prices. It is routinely evaluated by the Federal Statistical Office at the level of the German *Länder*. In the study provided here, model-based SAE techniques are applied to obtain estimates for the far-lower level German districts. As a first step, a standard Fay Herriot (FH) model was estimated, using regional indicators such as the population growth rate, the prevalence of rented housing and the price of building land as covariates. However, when this approach was discussed the approach with practitioners who work with rental price data on a daily basis, they decisively rejected the concept of a common relationship between rental prices and auxiliary information for all areas. They instead emphasized that the factors driving rental markets differ between different types of areas, such as rural and urban districts. They thus implicitly criticized the basic assumption of the Fay-Herriot model, i.e. a common relationship between the statistic of interest and the covariates for all areas, as being inappropriate. It was this criticism that motivated the extension of the standard small area model. The suggested approach of using a finite mixture of small area models may, however, prove appropriate in any application in which small area estimates for heterogeneous subentities are of interest.

This thesis is organized as follows: Chapter 2 and Chapter 3 are dedicated to introducing the two fields of basic theory that are relied upon in developing the proposal of a mixture of small area models. More specifically, Chapter 2 provides an overview of model-based SAE, including a general account of linear mixed models and a description of the basic unit- and area-level model as special cases. Chapter 3 introduces the basic theory of finite mixture models and presents relevant model definitions. This chapter also addresses the extension of a mixture model with a concomitant-variable submodel for the mixture weights. Furthermore, the choice of the number of components and the estimation of mixture models are discussed. In Chapter 4, the two fields introduced in Chapters 2 and 3 are brought together, and an estimator based on a finite mixture of small area models is presented. Corresponding to the structure of Chapter 2, the unit-level and the area-level model are introduced as special cases of a finite mixture of general linear mixed models. The topics of parameter estimation and prediction are also addressed. The proposed method is then evaluated in two model-based simulation studies discussed in Chapter 5. It is then applied to the problem of estimating rental prices at the district level in Germany. The thesis concludes with a summarizing evaluation of the proposed method and an outlook on future research.

Chapter 2

Model-based Small Area Statistics

2.1 Introduction

When a survey is conducted, there often is not only an interest in making statistical inferences about the entire population under study but also in obtaining reliable information for specified subentities. Then the following setting (see Münnich, Burgard and Vogt, 2013) is considered: A population $\mathcal{U}$ of size N is divided into m pairwise disjoint subpopulations $\mathcal{U}_i, i = 1, \ldots, m$. These subpopulations are called areas or domains depending on whether the disaggregation of the population is by region or by content.[1] A sample $\mathcal{S}$ of size n is drawn, with $\mathcal{S}_i = \mathcal{S} \cap \mathcal{U}_i$ designating the sample realized in $\mathcal{U}_i$ and n_i being the area-specific sample size. Now the aim is to simultaneously estimate a vector of m area-specific parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$, e.g. means $\boldsymbol{\mu}$.

When making inferences at the disaggregated level of the areas, researchers are commonly confronted with the problem of small sample sizes for the subentities. This is particularly true when the evaluation of a survey at a disaggregated level was not taken into account in the planning phase. But even if the sampling design and the area-specific sample sizes are thoroughly planned, practical or budgetary restrictions might lead to small subsamples for some or all areas (Jiang and Lahiri, 2006, p. 2). In this case, conventional direct estimators (which, by definition, only use the information from the area under consideration (Rao and Molina, 2015, p. 1)), might lead to large standard errors. Model-based SAE

[1] For the sake of brevity and simplicity, only the term *area* is used in the following discussion. In all theoretical considerations, the concepts are interchangeable.

methods are designed to produce reliable estimates, even for very small subsamples. Following Rao and Molina (2015, p.2), and corresponding to this aim, the eponymous *small area* is defined as an area with a subsample size that is too small to yield direct estimates of adequate precision. The strategy is to apply indirect techniques that make use of additional information such as, for example, sampled information from other areas. This additional information is included through a model. The intuition is that part of the (inter-area) variation of the statistic of interest can be explained by its relationship to a set of covariates valid for all areas under consideration. A model can, thus, be estimated using the data points from all subsamples. The specified relationship can then be employed to stabilize estimation. In the literature, this strategy is often referred to as "borrowing strength" (Ghosh and Rao, 1994).

Whereas classical direct estimators are generally design-unbiased, the unbiasedness of model-based estimators crucially depends on the validity of the assumed model. However, whenever direct estimators cannot be employed because some or all sample sizes are too small to yield estimates of adequate precision, the employment of model-based techniques is commonly regarded as best practice (Jiang and Lahiri (2006, p. 4), Pfeffermann (2002, p. 128), Rao and Molina (2015, p. 5), see Rao and Molina (2015, Chapter 3) for an account of competing design-based indirect approaches to SAE). As Pfeffermann (2002, p. 128) states, "SAE is widely recognized as one of the few problems in survey sampling where the use of models is often inevitable".

The standard approach is to estimate a linear mixed model, either at the level of the observation units or at the aggregated level of the areas. While the fixed part of the model establishes the above-mentioned constant relationship, which is used to stabilize the prediction, the random effect captures the variation between the areas that can not be explained by the covariates included in the model (see Jiang and Lahiri, 2006, p. 4). The standard area-level model was introduced by Fay and Herriot (1979). The unit-level model was first suggested by Battese, Harter and Fuller (1988). Both models are simple, special forms of the General Linear Mixed Model (GLMM). Therefore, the GLMM is introduced in Section 2.3. Then the standard area and the unit-level models are presented as special cases.

2.2 Literature Review

Small area estimation has atttracted much attention over the last decades, and standard methods are well established nowadays. An extensive overview is provided in the standard work by Rao (2003) and its second edition by Rao and

Molina (2015). Review papers focused on model-based SAE have been published by Jiang and Lahiri (2006), Pfeffermann (2002) and Pfeffermann (2013). Münnich et al. (2013) provide a German-language overview. A study that is thematically related to the application considered in this thesis has been published by Pereira and Coelho (2013), who estimated average house prices in Portugal by applying several small area estimators.

Concerning to the use of mixture models in SAE, a variety of suggestions, with different motives, have been made. Mixtures have been employed in order to relax restricting distributional assumptions or to model specific distributional shapes (see Chandra and Chambers, 2016; Elbers and van der Weide, 2014; Maiti, 2003), as well as in robust SAE (see Datta and Lahiri, 1995; Gershunskaya, 2010). Recently, Datta and Mandal (2015) proposed a mixture of a degenerate distribution localized at zero and the usual normal distribution for the random effects in an area-level mixed model. They, therewith, suggest a flexible SAE strategy, wherein random effects are only included for those areas for which the statistic of interest is not sufficiently well explained by the covariates included in the fixed part of the model. The proposal made in this thesis differs from all of these approaches with respect to its motivation and underlying intuition and, in particular, in the form in which mixtures are included in the framework. In this work, a mixture of mixed-effects regression models employed for the prediction of the statistic of interest in model-based SAE is considered.

With regard to clustering in SAE, Fabrizi, Montanari and Ranalli (2016) recently proposed latent class regression models for the classification of individual observations. Furthermore, clustering-based small area prediction, i.e. approaches that account for the existence of different subgroups of areas with subgroup-specific patterns when predicting the variable of interest, has also recently been considered: Torkashvand, Jafari Jozani and Torabi (2017) proposed area-level models in which the random effect distribution is allowed to vary between subgroups identified via hierarchical clustering based on the covariates. Maiti, Ren, Dass, Lim and Maier (2014) (see also Ren, 2011) considered an approach that allows both fixed and random effects to vary between clusters. The authors used the model-based clustering algorithm proposed by Booth, Casella and Hobert (2008) to partition areas into subgroups. Small area estimates are then obtained using cluster-specific Fay-Herriot models. In contrast, instead of a two-step procedure, this thesis considers an integrative approach that uses finite mixture models.

2.3 Linear Mixed Models

2.3.1 The Model

The GLMM (see Demidenko, 2004; Searle, Casella and McCulloch, 1992, p. 138-140) is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{v} + \boldsymbol{\epsilon} \tag{2.1}$$

with

$$\begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{pmatrix} \right), \tag{2.2}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of responses, $\boldsymbol{X}$ is an $n \times p$ design matrix of p known covariates and $\boldsymbol{\beta}$ is an $p \times 1$-vector of fixed effects. $\boldsymbol{v}$ denotes the $s \times 1$-vector of random effects and $\boldsymbol{U}$ is an $n \times s$ design matrix that defines in which form the random effects enter the model. $\boldsymbol{\epsilon}$ denotes the $n \times 1$-vector of error terms. $\boldsymbol{v}$ and $\boldsymbol{\epsilon}$ are assumed to be independently distributed with mean $\boldsymbol{0}$ and covariance matrices $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$, respectively.

Equations (2.1) and (2.2) imply the following (marginal) distribution of $\boldsymbol{y}$:

$$\begin{aligned} \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V}), \\ \boldsymbol{V} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T + \boldsymbol{\Sigma}. \end{aligned} \tag{2.3}$$

$\boldsymbol{D}$ and $\boldsymbol{\Sigma}$, and thus $\boldsymbol{V}$, are specified up to a vector of variance parameters $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_q)^T$. To make the dependence on $\boldsymbol{\vartheta}$ explicit, the notation $\boldsymbol{V}(\boldsymbol{\vartheta})$ is used whenever it clarifies the presentation.

The specification of the variance-covariance matrices $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$ defines how correlation structures in the data are captured by the model. The usual assumptions of independently and identically distributed error terms $\boldsymbol{\epsilon}$ are frequently made, such that $\boldsymbol{\Sigma} = \sigma_e^2 \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix.

The general formulation presented in (2.1) comprises a large family of different models. An important special case is the (multilevel or hierarchical) formulation for nested data, such as observations from repeated measurements in a longitudinal study or data comprising individual units grouped in contextual or regional aggregates as in SAE. Given such a data structure, mixed models are a device for taking correlations between observations from one individual or within clusters into account.

Consider m clusters of size N_i, $i = 1, \ldots, m$, with n_i sampled units in each cluster, where $n = \sum_{i=1}^{m} n_i$. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ denote the $n_i \times 1$ vector of observations for cluster i and $\boldsymbol{X}_i$ is the corresponding $n_i \times p$ matrix of covariates. Partitioning $\boldsymbol{y}, \boldsymbol{\epsilon}, \boldsymbol{v}$ and $\boldsymbol{X}$ into

$$\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_m^T)^T, \quad \boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_m^T)^T, \quad \boldsymbol{v} = (\boldsymbol{v}_1^T, \ldots, \boldsymbol{v}_m^T)^T \tag{2.4}$$

and

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix}$$

as well as specifying[2] $\boldsymbol{U} = \text{diag}(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m)$ and setting $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_m)$ as well as $\boldsymbol{D} = \text{diag}(\boldsymbol{G}, \ldots, \boldsymbol{G})$, yields the two-level Linear Mixed Model (LMM) (Demidenko (see 2004, p. 48-49) and Rao and Molina (2015, p. 108)), which can be decomposed into m submodels

$$\begin{gathered} \boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{v}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, m \\ \begin{pmatrix} \boldsymbol{v_i} \\ \boldsymbol{\epsilon_i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_i \end{pmatrix} \right). \end{gathered} \tag{2.5}$$

$\boldsymbol{v}_i$ is a r-vector, with $r = s/m$. This specification implies $\boldsymbol{V}_i = \boldsymbol{U}_i\boldsymbol{G}\boldsymbol{U}_i^T + \boldsymbol{R}_i$ and

$$\boldsymbol{V} = \text{diag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m) \tag{2.6}$$

i.e. a block diagonal covariance matrix, implying that observations from different clusters are uncorrelated. All small area models considered in the context of this thesis are special cases of (2.5).

2.3.2 Parameter Estimation

Parameter estimation for LMM requires both estimation of fixed effects $\boldsymbol{\beta}$ and of variance parameters $\boldsymbol{\vartheta}$. Commonly Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) estimation is employed to derive estimators for the model parameters. ML was first applied to LMM estimation by Hartley and Rao (1967). An extensive account is given by Demidenko (2004, Chapter 2) or Searle et al. (1992, Chapter 6).

[2] A simple but important special case is $\boldsymbol{U}_i = \boldsymbol{1}_{n_i}$, so that $\boldsymbol{U}$ is an $n \times m$ identity matrix, i.e. a common random intercept is assumed for observations within one cluster.

The likelihood function $L(\boldsymbol{\beta}, \boldsymbol{\vartheta})$ is obtained by considering the joint density of $\boldsymbol{y}$ as a function of the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$ given the data $\boldsymbol{y}$

$$L(\boldsymbol{\beta}, \boldsymbol{v}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right)}{(2\pi)^{n/2}|\boldsymbol{V}|^{1/2}}, \tag{2.7}$$

with corresponding log-likelihood

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\vartheta}) =& \log(L(\boldsymbol{\beta}, \boldsymbol{\vartheta})) \\ =& -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\log|\boldsymbol{V}| - \frac{n}{2}\log 2\pi. \end{aligned} \tag{2.8}$$

$|\boldsymbol{V}|$ denotes the determinant of $\boldsymbol{V}$. Maximizing (2.8) with respect to $\boldsymbol{\beta}$ by setting the partial derivative to zero yields the well-known Generalized Least Squares (GLS) estimator of $\boldsymbol{\beta}$:

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y}. \tag{2.9}$$

See McCulloch, Searle and Neuhaus (2008, pp. 163–164) and Searle et al. (1992, Appendix S.2.). It can be shown that (2.9) is the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$.

Calculating $\tilde{\boldsymbol{\beta}}$ requires knowledge of $\boldsymbol{V}(\boldsymbol{\vartheta})$, more specifically of $\boldsymbol{\vartheta}$, i.e. the $q \times 1$ vector of variance parameters it depends on. $\boldsymbol{\vartheta}$ is, however, usually unknown and has to be estimated from the data as well. To obtain a ML estimator of $\boldsymbol{\vartheta}$ the log-likelihood function has to be maximized with respect to the q elements in $\boldsymbol{\vartheta}$, too. Differentiating (2.8) and setting the partial derivative to zero gives

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\vartheta})}{\partial \vartheta_i} &= \frac{1}{2}\left\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} \frac{\partial \boldsymbol{V}}{\partial \vartheta_i} \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \operatorname{tr}\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \vartheta_i}\right)\right\} = 0, \\ & i = 1, \ldots, q. \end{aligned} \tag{2.10}$$

as first order condition for a maximum. Using $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y}$ in (2.10) results in the following condition:

$$\operatorname{tr}\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \vartheta_i}\right) = \boldsymbol{y}^T \boldsymbol{P} \frac{\partial \boldsymbol{V}}{\partial \vartheta_i} \boldsymbol{P}\boldsymbol{y}, \quad i = 1, \ldots, q, \tag{2.11}$$

with

$$\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \tag{2.12}$$

(see Jiang (2007, p. 10), McCulloch et al. (2008, p. 165), and Searle et al. (1992, Chapter 6.2)). Finding the ML estimator $\hat{\boldsymbol{\vartheta}}_{\text{ML}}$ requires solving (2.11) within

the parameter space (which is a restrictive requirement in the case of variance components) and second derivative tests to check whether the identified critical points indeed are maxima (see Searle et al., 1992, Chapter 6.3). The solution usually requires iterative procedures such as the Fisher-scoring algorithm (also denoted as Method of Scoring or Scoring algorithm) or the Newton-Raphson algorithm. See Rao and Molina (2015, p. 102) and Searle et al. (1992, p. 295) for the Fisher-scoring algorithm and Demidenko (2004, Chapter 2.8 – 2.15) or Searle et al. (1992, Chapter 8) for a comprehensive overview on numerical methods for computing ML estimates. Sometimes also the EM algorithm is employed to compute ML estimates. As this procedure is relied upon in the estimation of mixtures of mixed models in Chapter 4 it is described in detail in Section 2.3.5.

Once $\boldsymbol{\vartheta}$ is estimated from the data, an estimator for $\boldsymbol{\beta}$ is obtained calculating $\tilde{\boldsymbol{\beta}}$ from (2.9) substituting $\boldsymbol{V}(\boldsymbol{\vartheta})$ by an estimate $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\vartheta}}_{\text{ML}})$, i.e. $\hat{\boldsymbol{\beta}}_{\text{ML}} = \tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\vartheta}}_{\text{ML}})$ (Rao and Molina (2015, p. 102), Searle et al. (1992, Chapter 6.7)). Under the LMM as introduced in (2.5) (i.e. with normally distributed error terms), the estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\text{ML}}$, remains unbiased. For a more detailed discussion of properties see Demidenko (2004, Chapter 3.6).

It is known from more general results on ML estimation, that under suitable conditions the ML estimator for $\boldsymbol{\vartheta}$ is consistent and asymptotically normal distributed (see Demidenko (2004, Chapter 3.6)). The asymptotic covariance matrix $\overline{\boldsymbol{V}}$ of $\hat{\boldsymbol{\beta}}_{\text{ML}}$ and $\hat{\boldsymbol{\vartheta}}_{\text{ML}}$ is equal to the inverse of the Fisher information matrix $\boldsymbol{\mathcal{I}}$. Under certain regularity conditions, the Fisher information matrix is given by (see Jiang (2007, p. 11), Rao and Molina (2015, p. 102), Searle et al. (1992, Chapter 6.3))

$$\boldsymbol{\mathcal{I}} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\vartheta} \end{pmatrix} = -E \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\vartheta}^T} \\ \frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^T} \end{pmatrix}. \tag{2.13}$$

Assuming that $\boldsymbol{V}$ is twice continuously differentiable, the following expressions can be derived (Jiang (2007, p. 11), Searle et al. (1992, Chapter 6.3):

$$-E\left(\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right) = \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}, \tag{2.14}$$

$$-E\left(\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\vartheta}^T}\right) = -E\left(\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\beta}^T}\right) = \boldsymbol{0} \tag{2.15}$$

and $-E\left(\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}}\right) = \boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$, with $\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})$ being a $q \times q$-matrix with (i, j)-th element

given by

$$-E\left(\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\vartheta})}{\partial\vartheta_i\partial\vartheta_j}\right) = \frac{1}{2}\text{tr}(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \vartheta}_i \boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \vartheta}_j), \quad 1 \leq i,j \leq q. \tag{2.16}$$

The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{ML}}$ and $\hat{\boldsymbol{\vartheta}}_{\text{ML}}$, thus, has block-diagonal structure,

$$\overline{\boldsymbol{V}}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{\text{ML}}\\ \hat{\boldsymbol{\vartheta}}_{\text{ML}}\end{pmatrix} = \text{diag}((\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}, \boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta})^{-1}). \tag{2.17}$$

ML estimation of $\boldsymbol{\vartheta}$ does not take into account the degrees of freedom lost due to the estimation of $\boldsymbol{\beta}$ (Rao and Molina (2015, p. 102), Searle et al. (1992, Chapter 6.6)) and the result for the variance parameters depends on the fixed effects, which are sometimes considered as nuisance parameters and, thus, are of subordinate interest (Jiang, 2007, p. 12). Further, the estimator looses the property of consistency when the number of parameters in the model increases with the sample size (Jiang, 2007, p. 12, p. 40). An alternative estimation approach proposed by Patterson and Thompson (1971) and later Harville (1974) that overcomes these issues is REML estimation. Instead of the log-likelihood function for $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$, REML estimation maximizes the log-likelihood of the transformed data $\check{\boldsymbol{y}} = \boldsymbol{A}^T\boldsymbol{y}$, where $\boldsymbol{A}$ is any $n\times(n-p)$ matrix composed of $n-p$ linearly independent vectors $\boldsymbol{a}_1,\ldots,\boldsymbol{a}_{n-p}$ that fulfill $\boldsymbol{a}_i^T\boldsymbol{X} = \boldsymbol{0}^T$ for all $i = 1,\ldots,n-p$, with $p = \text{rank}(\boldsymbol{X})$ (Rao and Molina (2015, p. 102–103), Searle et al. (1992, Chapter 6.6)). The resulting distribution of the transformed data is $\check{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})$, i.e. it does not depend on $\boldsymbol{\beta}$. Thus, the fixed effects are "eliminated" (Jiang, 2007, p. 13) from the data.

The log-likelihood for $\check{\boldsymbol{y}}$, often denoted as "restricted" log-likelihood l_R, is given by

$$l_R(\boldsymbol{\vartheta}) = -\frac{1}{2}\check{\boldsymbol{y}}^T(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\check{\boldsymbol{y}} - \frac{1}{2}\log|\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A}| - \frac{(n-p)}{2}\log 2\pi. \tag{2.18}$$

Differentiating (2.18) with respect to the q elements of $\boldsymbol{\vartheta}$, setting the derivatives to zero and expressing the result in terms of $\boldsymbol{y}$ yields

$$\frac{\partial l_R(\boldsymbol{\vartheta})}{\partial\vartheta_i} = \frac{1}{2}\left\{\boldsymbol{y}^T\boldsymbol{P}\frac{\partial\boldsymbol{V}}{\partial\vartheta_i}\boldsymbol{P}\boldsymbol{y} - \text{tr}\left(\boldsymbol{P}\frac{\partial\boldsymbol{V}}{\partial\vartheta_i}\right)\right\} = 0, \qquad i = 1,\ldots,q, \tag{2.19}$$

where $\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}^T = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}$, as in (2.12) (Jiang, 2007, p. 13). The REML equations are, thus, given by

$$\text{tr}\left(\boldsymbol{P}\frac{\partial\boldsymbol{V}}{\partial\vartheta_i}\right) = \boldsymbol{y}^T\boldsymbol{P}\frac{\partial\boldsymbol{V}}{\partial\vartheta_i}\boldsymbol{P}\boldsymbol{y}, \quad i = 1,\ldots,q, \tag{2.20}$$

i.e. they differ from the respective ML equations (2.11) only insofar that $\boldsymbol{V}^{-1}$ on the left-hand side is replaced by $\boldsymbol{P}$. The REML estimator $\hat{\boldsymbol{\vartheta}}_{\text{REML}}$ is given by the solution of (2.20). As in the case of ML estimation, solving (2.20) requires iterative procedures. See Demidenko (2004, Chapter 2.14) and Rao and Molina (2015, p. 103) for details on the Fisher-Scoring algorithm for REML.

REML estimation does not provide an estimator for the fixed effect. As in ML estimation, a REML-estimator for $\boldsymbol{\beta}$ can, however, be obtained by substituting $\boldsymbol{V}(\boldsymbol{\vartheta})$ in (2.9) by an estimate $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\vartheta}}_{\text{REML}})$, so that $\hat{\boldsymbol{\beta}}_{\text{REML}} = \tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\vartheta}}_{\text{REML}})$ (Rao and Molina (2015, p. 103), Searle et al. (1992, Chapter 6.7)). For the LMM as defined in (2.5) the estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\text{REML}}$, again remains unbiased.

Under suitable conditions $\hat{\boldsymbol{\vartheta}}_{\text{REML}}$ is a consistent estimator for $\boldsymbol{\vartheta}$. It also is asymptotically normal. See Demidenko (2004, Chapter 3.6) and Jiang (2007, Chapter 1.8) for a more detailed discussion of statistical properties. If the number of fixed effects is fixed, the asymptotic covariance matrix is asymptotically equal to (2.17) (Rao and Molina, 2015, p. 103). Further, under the same condition, ML and REML are asymptotically equivalent (see Demidenko (2004, Chapter 3.6), Jiang (2007, p. 12, p. 40)). As Jiang (2007, p.40) points out, the true superiority of REML over ML estimation is, thus, revealed when considering a case where the number of fixed effects is large relative to the sample size. Additionally, it has been considered an advantage of REML that the estimates of the variance components are independent from the results obtained for the fixed effects and that, in the case of balanced data, REML solutions are equivalent to ANOVA estimators (McCulloch et al., 2008, Chapter 6.10). There, thus, has evolved a certain preference for REML, which is also the default setting in standard R-packages for estimating Mixed Models, `lme4` (Bates, Mächler, Bolker and Walker, 2015) and `nlme` (Pinheiro, Bates, DebRoy, Sarkar and R Core Team, 2017), and the standard package for SAE, `sae` (Molina and Marhuenda, 2015).

2.3.3 Mixed Model Prediction

In SAE and in many other applications of mixed models, the prediction from the mixed model, i. e. one that involves both the fixed effect and the realized value of the random effect, is of interest (Harville, 1976; Harville and Jeske, 1992; Henderson, 1975; Jiang and Lahiri, 2006; Rao and Molina, 2015). The respective general linear combination of both fixed and random effects of the form $\eta = \boldsymbol{l}^T\boldsymbol{\beta} + \boldsymbol{m}^T\boldsymbol{v}$, where $\boldsymbol{l}$ and $\boldsymbol{m}$ are given vectors of constants, (Jiang and Lahiri (2006, Chapter 3.2), Rao and Molina (2015, Chapter 5.2)) is sometimes referred to as "mixed effect" (Jiang and Lahiri, 2006, p. 12).

Obviously, mixed model prediction requires assigning values to the (unobservable)

random effects, i.e. estimating the realized (but unobservable) value of the random variable $\boldsymbol{v}$ given the data (McCulloch et al., 2008, Chapter 1.7). Although this implies estimating a "thing [that] has already occurred" (Robinson, 1991, p. 28), this task has evolved to be commonly referred to as *prediction* of random effects (Robinson, 1991). As usual in the respective literature, in the following account, it is first assumed that all parameters of the model are known. It then is easy to show that the Best Predictor (BP) for $\boldsymbol{v}$, i.e. the one that minimizes the Mean Square Error (MSE) of prediction $\text{MSE}(\tilde{\boldsymbol{v}}) = E(\tilde{\boldsymbol{v}} - \boldsymbol{v})^2$ with $\tilde{\boldsymbol{v}}$ denoting a predictor, is the conditional expectation of $\boldsymbol{v}$ given the observed data $\boldsymbol{y}$:

$$\tilde{\boldsymbol{v}} = BP(\boldsymbol{v}) = E(\boldsymbol{v}|\boldsymbol{y}). \tag{2.21}$$

For the LMM with normally distributed errors as defined in (2.5), $E(\boldsymbol{v}|\boldsymbol{y})$ and therewith the BP $\tilde{\boldsymbol{v}}$ can straightforwardly be deduced from the multivariate normal joint distribution of $\boldsymbol{v}$ and $\boldsymbol{y}$ and respective standard results on the conditional distribution of $\boldsymbol{v}$ given $\boldsymbol{y}$ (see Searle et al., 1992, Chapter 7.3 and appendix S.3.) as

$$\tilde{\boldsymbol{v}} = E(\boldsymbol{v}|\boldsymbol{y}) = \boldsymbol{D}\boldsymbol{U}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{2.22}$$

Note that the same result is obtained when the Best Linear Predictor (BLP) is of interest, i.e. when the MSE is minimized for a *linear* predictor of the form $\tilde{\boldsymbol{v}} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{y}$ where $\boldsymbol{a}$ and $\boldsymbol{B}$ denote some vector and matrix, respectively. The derivation then does not require the assumption of normality (see Searle et al., 1992, Chapter 7.3). For the linear combination $\eta = \boldsymbol{l}^T\boldsymbol{\beta} + \boldsymbol{m}^T\boldsymbol{v}$, the BP under normality is derived, correspondingly, as $\tilde{\eta} = E(\eta|\boldsymbol{y}) = \boldsymbol{l}^T\boldsymbol{\beta} + \boldsymbol{m}^T\tilde{\boldsymbol{v}}$. Again this estimator is also the BLP without requiring any distributional assumption (Rao and Molina, 2015, Chapter 5.2.1).

Of course, model parameters are usually unknown. If the fixed effects have to be inferred from the data, but variance parameters $\boldsymbol{\vartheta}$ are still known, $\boldsymbol{\beta}$ is commonly replaced by its GLS estimator $\tilde{\boldsymbol{\beta}}$, ie. the BLUE of $\boldsymbol{\beta}$ as given in (2.9), such that a predictor for η is given by

$$\tilde{\eta}^{BLUP} = \boldsymbol{l}^T\tilde{\boldsymbol{\beta}} + \boldsymbol{m}^T\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}). \tag{2.23}$$

This estimator was proposed by Henderson (1950) and is commonly referred to as the Best Linear Unbiased Predictor (BLUP) for η, a label first used by Goldberger (1962). The famous acronym BLUP was later coined by Henderson (1973). Note that

$$\tilde{\boldsymbol{v}}^{BLUP} = \boldsymbol{D}\boldsymbol{U}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}), \tag{2.24}$$

i.e. $\tilde{\boldsymbol{v}}$ with $\boldsymbol{\beta}$ replaced by $\tilde{\boldsymbol{\beta}}$, also is denoted the BLUP for $\boldsymbol{v}$ (Demidenko, 2004, Chapter 3.7).

The BLUP given in (2.23) is a *linear* function of $\boldsymbol{y}$. It also is *best* in the sense that it minimizes the MSE of prediction and *unbiased* in the sense that $E(\hat{\eta}^{BLUP}) = E(\eta)$. Note that the meaning of these concepts thus differs from their definitions in the context of estimation of a fixed value: As now a random variable is predicted, the MSE is minimized instead of the variance and the relevant notion of unbiasedness is that the expected value of the predictor equals the expected value of the random variable instead of the fixed quantity to be estimated (Robinson (1991), Searle et al. (see 1992, Chapter 7.2)). A proof of these properties is given in Henderson (1963) or Rao and Molina (2015, Chapter 5.6.1) and Searle et al. (1992, Chapter 7.4).

The BLUP was first derived by Henderson. In search of what he called "joint maximum likelihood estimates", Henderson (1950) simultaneously maximized the joint density of $\boldsymbol{y}$ and $\boldsymbol{v}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{v}$. Setting the partial derivatives to zero results in the following system of equations (Henderson, 1950; Henderson, Kempthorne, Searle and von Krosigk, 1959)

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{U} \\ \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{U} + \boldsymbol{D}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{v}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \\ \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \end{bmatrix}. \tag{2.25}$$

They have come to be known as Henderson's *Mixed Model Equations*. The solutions of (2.25) correspond to (2.9) and (2.24). These equations are often computationally more economic than the standard BLUE and BLUP equations because inversion of $\boldsymbol{\Sigma}$ and $\boldsymbol{D}$ is often easier than of $\boldsymbol{V}$ (Henderson et al. (1959), Searle et al. (1992, Chapter 7.6)).

As Rao points out, the above considerations can straightforwardly be extended to the case where two or more linear combinations are to be estimated simultaneously, i.e. where the estimation of $\boldsymbol{\eta} = \boldsymbol{L}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{v}$ is of interest (Rao and Molina (2015, p. 100), see Searle et al. (1992, Chapter 7.4) for a corresponding presentation of the topic).

So far it was assumed that the variance parameters $\boldsymbol{\vartheta}$ are known. In practice, however, they usually have to be estimated from the data as well and $\boldsymbol{\vartheta}$ in (2.23) is replaced by an estimator $\hat{\boldsymbol{\vartheta}}$. The resulting (two-stage) estimator (Kackar and Harville (1981, p. 1256), Rao and Molina (2015, p. 101)) $\hat{\eta}$ is referred to as empirical BLUP or Empirical Best Linear Unbiased Predictor (EBLUP). Kackar and Harville (1981) showed that the EBLUP remains an unbiased estimator of η under the conditions that $\hat{\boldsymbol{\vartheta}}$ is a translation-invariant estimator of $\boldsymbol{\vartheta}$ and an even function of the data, $\boldsymbol{v}$ and $\boldsymbol{\epsilon}$ are both distributed symmetrically around $\boldsymbol{0}$, and $E(\hat{\eta})$ exists. They further showed that standard estimators of $\boldsymbol{\vartheta}$, including

both ML and REML estimators, are even and translation-invariant (Kackar and Harville, 1981).

2.3.4 MSE Estimation

To assess the accuracy of the EBLUP $\hat{\eta}$, the MSE of prediction, i.e. $\text{MSE}(\hat{\eta}) = E(\hat{\eta} - \eta)^2$, is considered. This task is highly complex, particularly because it requires assessing the variability caused by the estimation of $\boldsymbol{\vartheta}$.

Ignoring the added variation due to the estimation of variance parameters, the MSE of the EBLUP is sometimes approximated by using the expression for the MSE of the BLUP, $\text{MSE}(\tilde{\eta}^{\text{BLUP}})$, which is known to be (see Henderson (1975), Rao and Molina (2015, Chapter 5.2.2))

$$\text{MSE}(\tilde{\eta}^{\text{BLUP}}) = g_1(\boldsymbol{\vartheta}) + g_2(\boldsymbol{\vartheta}), \tag{2.26}$$

with

$$g_1(\boldsymbol{\vartheta}) = \boldsymbol{m}^T(\boldsymbol{D} - \boldsymbol{D}\boldsymbol{U}^T\boldsymbol{V}^{-1}\boldsymbol{U}\boldsymbol{D})\boldsymbol{m} \tag{2.27}$$

$$g_2(\boldsymbol{\vartheta}) = (\boldsymbol{l} - \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{U}\boldsymbol{D}\boldsymbol{m})^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{l} - \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{U}\boldsymbol{D}\boldsymbol{m})^T, \tag{2.28}$$

and replacing $\boldsymbol{\vartheta}$ by its estimate $\hat{\boldsymbol{\vartheta}}$. This naive estimator might, however, underestimate the MSE to a considerable degree. Kackar and Harville (1984) showed that, for normally distributed error terms and provided $\hat{\boldsymbol{\vartheta}}$ is a translation-invariant estimator, the MSE of the EBLUP can be decomposed as follows:

$$\text{MSE}(\hat{\eta}) = \text{MSE}(\tilde{\eta}^{\text{BLUP}}) + E(\hat{\eta} - \tilde{\eta}^{\text{BLUP}})^2. \tag{2.29}$$

Using (2.26) as an estimator, thus, implies neglecting the second term of (2.29), which represents the contribution to the MSE due to the estimation of $\boldsymbol{\vartheta}$. This underestimation, which might be significant if the variation of $\hat{\boldsymbol{\vartheta}}$ is large and $\hat{\eta}$ varies strongly with $\boldsymbol{\vartheta}$ (Rao and Molina, 2015, Chapter 5.2.5), might not be acceptable in practice.

With very few exemptions, $E(\hat{\eta} - \tilde{\eta}^{\text{BLUP}})^2$ is, however, intractable (Rao and Molina, 2015, Chapter 5.2.5) and an approximation

$$g_3(\boldsymbol{\vartheta}) \approx E(\hat{\eta} - \tilde{\eta}^{\text{BLUP}})^2 \tag{2.30}$$

is required. A path-breaking contribution to this research task was made by Prasad and Rao (1990). They derived an approximation for the FH-model (see Section 2.4) and the nested error regression model (see Section 2.5) and provided results on its accuracy. Datta and Lahiri (2000) presented more general

results for the LMM with block-diagonal covariance structure (as defined in (2.5)) with variance components estimated by ML and REML. Das, Jiang and Rao (2004) further extended this work by considering a general GLMM as in (2.1). See Das et al. (2004) or Rao and Molina (2015, Chapter 5.2.5) for the respective expression for $g_3(\boldsymbol{\vartheta})$. Results for the special cases of the FH- and the nested error regression models are given in Section 2.4 and 2.5, respectively. All three approximations neglect terms of order $o(m^{-1})$, such that a second-order unbiased estimator for $\text{MSE}(\hat{\eta})$ is given by

$$\begin{aligned}\text{MSE}(\hat{\eta}) &\approx \text{MSE}(\tilde{\eta}^{\text{BLUP}}) + g_3(\boldsymbol{\vartheta}) \\ &\approx g_1(\boldsymbol{\vartheta}) + g_2(\boldsymbol{\vartheta}) + g_3(\boldsymbol{\vartheta}).\end{aligned} \tag{2.31}$$

In practice, an estimator $\widehat{\text{MSE}}(\hat{\eta})$ for $\text{MSE}(\hat{\eta})$ has to be obtained. $g_2(\boldsymbol{\vartheta})$ and $g_3(\boldsymbol{\vartheta})$ can generally be approximated by $g_2(\hat{\boldsymbol{\vartheta}})$ and $g_3(\hat{\boldsymbol{\vartheta}})$, respectively. Further, when variance components are estimated by REML estimation,

$$g_1(\boldsymbol{\vartheta}) \approx E(g_1(\hat{\boldsymbol{\vartheta}}) + g_3(\hat{\boldsymbol{\vartheta}})). \tag{2.32}$$

An estimator for $\text{MSE}(\hat{\eta})$ with bias of order $o(m^{-1})$ is then derived as

$$\widehat{\text{MSE}}_{\text{REML}}(\hat{\eta}) \approx g_1(\hat{\boldsymbol{\vartheta}}) + g_2(\hat{\boldsymbol{\vartheta}}) + 2g_3(\hat{\boldsymbol{\vartheta}}). \tag{2.33}$$

For the ML estimator of variance components, the following approximation holds:

$$\widehat{\text{MSE}}_{\text{ML}}(\hat{\eta}) \approx g_1(\hat{\boldsymbol{\vartheta}}) - \boldsymbol{b}^T(\hat{\boldsymbol{\vartheta}};\boldsymbol{\vartheta})\frac{\partial g_1(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} + g_2(\hat{\boldsymbol{\vartheta}}) + 2g_3(\hat{\boldsymbol{\vartheta}}), \tag{2.34}$$

where $\boldsymbol{b}^T(\hat{\boldsymbol{\vartheta}};\boldsymbol{\vartheta})$ is an approximation of the bias of $\hat{\boldsymbol{\vartheta}}$. An expression of the additional bias correction term for the special case of a standard area-level model is given in Sections 2.4. See Rao and Molina (2015, Chapter 5.2.6) and Das et al. (2004) for details.

2.3.5 The EM Algorithm for Mixed Models

In Section 2.3.2 it was stated that ML and REML estimates for the LMM are calculated employing numerical procedures such as the Fisher scoring or Newton-Raphson algorithm. A third algorithm commonly employed is the EM algorithm, a multi-purpose estimation algorithm introduced generally in Appendix A.1 (see Demidenko, 2004, Chapter 2.8 for a comparative overview of the three approaches). The EM algorithm was applied early for maximum likelihood inference for linear mixed models (Jennrich and Schluchter, 1986; Laird, Lange

and Stram, 1987; Laird and Ware, 1982; Lindstrom and Bates, 1988). See Searle et al. (1992, Chapter 8.3) and McLachlan and Krishnan (2008, Chapter 5.9) for an overview. As this algorithm is relied upon and extended in the context of estimation of a mixture of mixed models, the calculation of ML estimates for the LMM via the EM algorithm is briefly presented here (for calculation of REML estimates see e.g. Searle et al. (1992, Chapter 8.3)). As only the special case of a LMM with block-diagonal covariance-structure (as defined in (2.5)) and independently distributed error terms, i.e. $\boldsymbol{R}_i = \sigma_e^2 \boldsymbol{I}_{n_i}$, is of interest in the context of this thesis, the presentation is restricted to a model with this simplified covariance structure.

Consistent with the approach described in Appendix A.1, for estimating the LMM via the EM-algorithm, the model is interpreted as a missing data situation. More specifically, the random effects $\boldsymbol{v}_i, i = 1 \ldots, m$ are treated as missing values, exploiting the fact that estimation of model parameters would be trivial if they were observed (Searle et al. (1992, Chapter 8.3)).

The joint distribution of the complete-data vector of area i, $(\boldsymbol{y}_i^T, \boldsymbol{v}_i^T)^T$ is the multivariate normal density

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{v}_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{X_i\beta} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V}_i & \boldsymbol{U}_i\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{U}_i^T & \boldsymbol{G} \end{pmatrix} \right), \tag{2.35}$$

where $\boldsymbol{V}_i = \boldsymbol{U}_i \boldsymbol{G} \boldsymbol{U}_i^T + \sigma_e^2 \boldsymbol{I}_{n_i}$. Denoting the variance covariance matrix by $\boldsymbol{\Sigma}_i$, defining the vector

$$\boldsymbol{d}_i = \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{X_i\beta} \\ \boldsymbol{v}_i - \boldsymbol{0} \end{pmatrix}, \tag{2.36}$$

and using the fact that for any $i \neq j$, $\text{Cov}(\boldsymbol{y}_i, \boldsymbol{y}_j^T) = 0$, $\text{Cov}(\boldsymbol{v}_i, \boldsymbol{v}_j^T) = 0$ and $\text{Cov}(\boldsymbol{y}_i, \boldsymbol{v}_j^T) = 0$, the log-likelihood based of the complete-data vector $(\boldsymbol{y}^T, \boldsymbol{v}^T)^T = (\boldsymbol{y}_i^T, \ldots, \boldsymbol{y}_m^T, \boldsymbol{v}_i^T, \ldots, \boldsymbol{v}_m^T)^T$ is accordingly given by

$$l_c(\boldsymbol{\psi}) = -\frac{1}{2} \sum_{i=1}^{m} \left(\boldsymbol{d}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{d}_i + \log |\boldsymbol{\Sigma}_i| + (n_i + s) \log(2\pi) \right). \tag{2.37}$$

where s is the length of the random effect-vector $\boldsymbol{v}_i$. $\boldsymbol{\psi}$ is the vector of unknown parameters comprising $\boldsymbol{\beta}, \sigma_e^2$ and the vector of variance parameters $\boldsymbol{\vartheta}_v$, $\boldsymbol{G}$ depends on, i.e. $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \sigma_e^2, \boldsymbol{\vartheta}_v^T)^T$.

As the random effects are, of course, not observable so that (2.37) cannot be formed, the EM algorithm alternatively works on the expectation of $l_c(\boldsymbol{\psi})$, i.e. $Q(\boldsymbol{\psi}; \widehat{\boldsymbol{\psi}}^{(t-1)}) = E_{\widehat{\boldsymbol{\psi}}^{(t-1)}} [l_c(\boldsymbol{\psi})]$. Deriving Q requires the conditional distribution of

$\boldsymbol{v}_i$ given $\boldsymbol{y}_i$, which from standard multivariate normal theory (see e.g. Fahrmeir, Kneib, Lang and Marx, 2013, Theorem B.6 on p. 649) is known to be

$$\boldsymbol{v}_i|\boldsymbol{y}_i \sim \mathcal{N}\left(\boldsymbol{G}\boldsymbol{U}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \boldsymbol{G} - \boldsymbol{G}\boldsymbol{U}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{U}_i\boldsymbol{G}\right). \tag{2.38}$$

Using $\boldsymbol{V}_i = \boldsymbol{U}_i\boldsymbol{G}\boldsymbol{U}_i^T + \sigma_e^2\boldsymbol{I}_{n_i}$ it can also be expressed as (see McLachlan and Krishnan, 2008, Chapter 5.9)

$$\boldsymbol{v}_i|\boldsymbol{y}_i \sim \mathcal{N}\left((\boldsymbol{U}_i^T\boldsymbol{U} + \sigma_e^2\boldsymbol{G}^{-1})^{-1}\boldsymbol{U}_i^T(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), ((\sigma_e^2)^{-1}\boldsymbol{U}_i^T\boldsymbol{U}_i + \boldsymbol{G}^{-1})^{-1}\right). \tag{2.39}$$

Using (2.37)–(2.39), the EM algorithm is performed applying the following steps (see Appendix A.1 for more details and a general description of the algorithm):

- *Specification of starting values* $\widehat{\boldsymbol{\psi}}^{(0)}$
- *E-step in iteration (t)*
  Deriving $Q(\boldsymbol{\psi}; \widehat{\boldsymbol{\psi}}^{(t-1)})$ particularly requires deriving the conditional expectation of the sufficient statistics $\boldsymbol{v}_i$ and $\boldsymbol{v}_i\boldsymbol{v}_i^T$ given $\boldsymbol{y}_i$ (see e.g. de Leeuw and Meijer, 2008, Appendix 1.D.) using estimates $\hat{\sigma}_e^{(t-1)^2}$, $\widehat{\boldsymbol{G}}^{(t-1)} = \boldsymbol{G}(\hat{\boldsymbol{\vartheta}}_v^{(t-1)})$ and $\hat{\boldsymbol{\beta}}^{(t-1)}$ of σ_e^2, $\boldsymbol{G}$ and $\boldsymbol{\beta}$ obtained in the last iteration:

$$\begin{aligned}\hat{\boldsymbol{s}}_{1i}^{(t-1)} &= E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(\boldsymbol{v}_i|\boldsymbol{y}_i) \\ &= (\boldsymbol{U}_i^T\boldsymbol{U}_i + \hat{\sigma}_e^{2^{(t-1)}}\widehat{\boldsymbol{G}}^{(t-1)^{-1}})^{-1}\boldsymbol{U}_i^T(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(t-1)}),\end{aligned} \tag{2.40}$$

  and

$$\begin{aligned}\widehat{\boldsymbol{S}}_{2i}^{(t-1)} &= E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(\boldsymbol{v}_i\boldsymbol{v}_i^T|\boldsymbol{y}_i) \\ &= \mathrm{Cov}_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(\boldsymbol{v}_i|\boldsymbol{y}_i) + \hat{\boldsymbol{s}}_{1i}^{(t-1)}\hat{\boldsymbol{s}}_{1i}^{(t-1)^T} \\ &= ((\sigma_e^{2^{(t-1)}})^{-1}\boldsymbol{U}_i^T\boldsymbol{U}_i + \widehat{\boldsymbol{G}}^{(t-1)^{-1}})^{-1} + \widehat{\boldsymbol{s}}_{1i}^{(t-1)}\widehat{\boldsymbol{s}}_{1i}^{(t-1)^T}.\end{aligned} \tag{2.41}$$

  Here and in the following, $E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}$ denotes expectation parameterized by $\boldsymbol{\psi}^{(t-1)}$, where $\widehat{\boldsymbol{\psi}}^{(t-1)}$ is the vector of estimates obtained in the last iteration step $(t-1)$.
- *M-step*
  An updated estimate of σ_e^2, $\boldsymbol{G}$ and $\boldsymbol{\beta}$ is derived by maximizing $Q(\boldsymbol{\psi}; \widehat{\boldsymbol{\psi}}^{(t-1)})$, i.e. the maximum likelihood estimates of the complete-data log-likelihood

(2.37) are calculated substituting the (unknown) sufficient statistics with their conditional expected values obtained in the E-step. Thus,

$$\widehat{\boldsymbol{\beta}}^{(t)} = (\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{X}_i)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T (\boldsymbol{y}_i - \boldsymbol{U}_i \hat{\boldsymbol{s}}_{1i}^{(t-1)}), \tag{2.42}$$

$$\widehat{\boldsymbol{G}}^{(t)} = \frac{1}{m} \sum_{i=1}^{m} \widehat{\boldsymbol{S}}_{2i}^{(t-1)}, \tag{2.43}$$

and

$$\hat{\sigma}_e^{2(t)} = \frac{1}{n} \sum_{i=1}^{m} E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(\boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i | \boldsymbol{y}_i). \tag{2.44}$$

With

$$\hat{\boldsymbol{\epsilon}}_i^{(t-1)} = E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(\boldsymbol{\epsilon}_i | \boldsymbol{y}_i) = \boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{U}_i \hat{\boldsymbol{s}}_{1i}^{(t-1)}, \tag{2.45}$$

and

$$\begin{aligned} \mathrm{Cov}_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(\boldsymbol{\epsilon}_i | \boldsymbol{y}_i) &= \mathrm{Cov}_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(\boldsymbol{U}_i \boldsymbol{v}_i | \boldsymbol{y}_i) \\ &= \boldsymbol{U}_i ((\hat{\sigma}_e^{2^{(t-1)}})^{-1} \boldsymbol{U}_i^T \boldsymbol{U}_i + \widehat{\boldsymbol{G}}^{(t-1)^{-1}})^{-1} \boldsymbol{U}_i^T \end{aligned} \tag{2.46}$$

using standard results on the expected value of quadratic forms (see e.g. McCulloch et al., 2008, Appendix S.1) this yields

$$\begin{aligned} \hat{\sigma}_e^{2^{(t)}} = \frac{1}{n} \sum_{i=1}^{m} [& \hat{\boldsymbol{\epsilon}}_i^{(t-1)^T} \hat{\boldsymbol{\epsilon}}_i^{(t-1)} \\ & + \mathrm{tr} \left(\boldsymbol{U}_i^T \boldsymbol{U}_i ((\hat{\sigma}_e^{2^{(t-1)}})^{-1} \boldsymbol{U}_i^T \boldsymbol{U}_i + \widehat{\boldsymbol{G}}^{(t-1)^{-1}})^{-1} \right)]. \end{aligned} \tag{2.47}$$

See McLachlan and Krishnan (2008, Chapter 5.9).

- *Termination*
  Both steps are repeated until convergence (see Appendix A.1).

2.4 The Fay-Herriot Model

Assume that the true area mean μ_i is appropriately modelled by a linear model $\mu_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + v_i$ (linking model), where v_i is an area-specific random effect, with

$v_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_v^2)$. $\boldsymbol{x}_i^{\mathrm{T}}$ denotes the vector of auxiliary information for area i and $\boldsymbol{\beta}$ is a vector of corresponding regression coefficients. Furthermore, direct estimates $\hat{\mu}_i^{\mathrm{Dir}}$, calculated from the sample realized in area $i, i = 1, \ldots, m$, are assumed to be related to μ_i in the form $\hat{\mu}_i^{\mathrm{Dir}} = \mu_i + e_i$ (sampling model), where e_i is the sampling error of the direct estimate, $e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_{e,i}^2)$, with known design variance $\sigma_{e,i}^2$.

Combining these assumptions yields

$$\begin{aligned} \hat{\mu}_i^{\mathrm{Dir}} &= \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + v_i + e_i \qquad \text{for} \quad i = 1, \ldots, m \\ v_i &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_v^2) \\ e_i &\overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_{e,i}^2). \end{aligned} \tag{2.48}$$

It is assumed that v_i and e_i are independent.

(2.48) was introduced by FAY and HERRIOT (1979) in the context of estimating per capita income for small areas in the USA and is, therefore, commonly referred to as Fay-Herriot (FH) model. Since then, it has become a standard model for small area estimation.

Note that (2.48) is a simple special case of the LMM with block-diagonal covariance structure as defined in (2.5), where

$$\boldsymbol{y_i} = \hat{\mu}_i^{\mathrm{Dir}}, \quad \boldsymbol{X}_i = \boldsymbol{x}_i^T, \quad \boldsymbol{U}_i = 1, \tag{2.49}$$

$$\boldsymbol{v}_i = v_i, \qquad \boldsymbol{\epsilon}_i = e_i, \qquad \boldsymbol{G} = \sigma_v^2, \qquad \boldsymbol{R}_i = \sigma_e^2,$$

and thus

$$\boldsymbol{V}_i = \sigma_{e,i}^2 + \sigma_v^2. \tag{2.50}$$

The following brief account of parameter estimation, prediction and MSE estimation for the FH model can thus heavily draw on the details given for the general model in the preceding sections 2.3.2 – 2.3.4.

Along the lines of RAO and MOLINA (2015, Chapter 6.1.1), the BLUP for the parameter of interest μ under this model can easily be deduced from the general expressions for the BLUP of η and the BLUE for $\boldsymbol{\beta}$ by making the corresponding substitutions in (2.23) and (2.9). Noting that μ_i is a special case of the general mixed effect $\eta_i = \boldsymbol{l}_i^T\boldsymbol{\beta} + \boldsymbol{m}_i^T\boldsymbol{v}_i$, where $\boldsymbol{l}_i = \boldsymbol{x}_i$ and $\boldsymbol{m}_i = 1$ this yields

$$\begin{aligned} \tilde{\mu}_i^{\mathrm{FH}} &= \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}} + \tilde{v}_i^{\mathrm{BLUP}} \\ &= \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}} + \gamma_i(\hat{\mu}_i^{\mathrm{Dir}} - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) \\ &= \gamma_i\hat{\mu}_i^{\mathrm{Dir}} + (1 - \gamma_i)\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}} \end{aligned} \tag{2.51}$$

with

$$\gamma_i = \frac{\sigma_v^2}{\sigma_{e,i}^2 + \sigma_v^2}. \tag{2.52}$$

and

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_{e,i}^2 + \sigma_v^2} \right)^{-1} \left(\sum_{i=1}^{m} \frac{\boldsymbol{x}_i \hat{\mu}_i^{\text{Dir}}}{\sigma_{e,i}^2 + \sigma_v^2} \right). \tag{2.53}$$

(2.51) is called the Fay Herriot (FH) estimator for the parameter of interest. Note that it can be expressed as a composite estimator of the synthetic estimator $\boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}$, obtained from the fixed part of the model, and the direct estimator $\hat{\mu}_i^{\text{Dir}}$. Weights are given by the area-specific shrinkage factor γ_i, that sets the model variance σ_v^2 in relation to the total variance $\sigma_{e,i}^2 + \sigma_v^2$. If the model variance is small compared to the design variance, γ_i is close to zero and the synthetic estimator dominates. Intuitively, γ_i can be understood as a relative measure of confidence in the model- and the design-based estimator (see Rao and Molina, 2015, Chapter 6.1.1).

As in Section 2.3.3, the EBLUP is obtained from the BLUP by replacing the typically unknown variance components by an estimate $\hat{\boldsymbol{\vartheta}}$. In the case of the FH model $\boldsymbol{\vartheta} = \sigma_v^2$ and the EBLUP is given by

$$\hat{\mu}_i^{\text{FH}} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \hat{\gamma}_i (\hat{\mu}_i^{\text{Dir}} - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}), \tag{2.54}$$

with

$$\hat{\gamma}_i = \frac{\hat{\sigma}_v^2}{\sigma_{e,i}^2 + \hat{\sigma}_v^2} \tag{2.55}$$

and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_v^2)$.

As in the general case (see Section 2.3.2), the variance parameter σ_v^2 can be estimated via ML or REML estimation. Again, iterative procedures are applied for computation. See Rao and Molina (2015, Chapter 6.1.2) for details on the Fisher-scoring algorithm for the special case of the FH model. ML and REML estimates can also be derived employing the EM algorithm (see Section 2.3.5). Alternatively, σ_v^2 can also be estimated by the method of moments (Fay and Herriot (1979), Prasad and Rao (1990)). This was the estimation approach, Fay and Herriot (1979) originally employed in their seminal paper introducing the FH model. As described in Section 2.3.3, the asymptotic variances of ML and REML estimators are obtained as the inverse of the Fisher-information $(\mathcal{I}(\sigma_v^2))^{-1}$.

It is given by

$$\overline{V}(\hat{\sigma}^2_{v,\mathrm{ML}}) = \overline{V}(\hat{\sigma}^2_{v,\mathrm{REML}}) = 2\left(\sum_{i=1}^{m}\frac{1}{(\sigma_v^2+\sigma_e^2)^2}\right)^{-1}. \tag{2.56}$$

For results for the asymptotic variances of moment estimators see Prasad and Rao (1990) and Datta, Rao and Smith (2005).

Regarding the MSE of the FH estimator, $\mathrm{MSE}(\hat{\mu}_i^{\mathrm{FH}}) = E(\hat{\mu}_i^{\mathrm{FH}} - \mu_i)^2$, general results from Section 2.3.4 can be relied upon. In accordance with (2.31), the MSE of $\hat{\mu}_i^{\mathrm{FH}}$ is approximated by

$$\mathrm{MSE}(\hat{\mu}_i^{FH}(\sigma_v^2)) = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2). \tag{2.57}$$

g_1 and g_2 can be obtained from the expressions (2.27) and (2.28), respectively. It is

$$g_{1i}(\sigma_v^2) = \frac{\sigma_v^2\sigma_{e,i}^2}{\sigma_v^2+\sigma_{e,i}^2}, \tag{2.58}$$

$$g_{2i}(\sigma_v^2) = \left(\frac{\sigma_{e,i}^2}{\sigma_v^2+\sigma_{e,i}^2}\right)^2 \boldsymbol{x}_i^T\left(\sum_{j=1}^{m}\frac{\boldsymbol{x}_j\boldsymbol{x}_j^T}{(\sigma_{e,j}^2+\sigma_v^2)}\right)^{-1}\boldsymbol{x}_i. \tag{2.59}$$

Further, as discussed in Section 2.3.5, based on work by Prasad and Rao (1990), Datta and Lahiri (2000) provided a second order approximation of g_3 for σ_v^2 estimated by ML or REML given by

$$g_{3i}(\sigma_v^2) = \frac{\sigma_{e,i}^4}{(\sigma_v^2+\sigma_{e,i}^2)^3}\frac{2}{\sum_{j=1}^{m}\frac{1}{(\sigma_v^2+\sigma_{e,j}^2)^2}}. \tag{2.60}$$

Finally, in line with the general results, an estimator for $\mathrm{MSE}(\hat{\mu}_i^{FH})$ with bias of order $o(m^{-1})$ can be obtained by estimating $g_{2i}(\sigma_v^2)$ and $g_{3i}(\sigma_v^2)$ by $g_{2i}(\hat{\sigma}_v^2)$, respectively. For σ_v^2 estimated by REML $g_{3i}(\hat{\sigma}_v^2)$, furthermore $g_{1i}(\sigma_v^2)$ can be estimated by $g_{1i}(\hat{\sigma}_v^2) + g_{3i}(\hat{\sigma}_v^2)$ so that $\widehat{\mathrm{MSE}}_{\mathrm{REML}}(\hat{\mu}_i^{FH}) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2)$. If σ_v^2 is estimated by ML, $\widehat{\mathrm{MSE}}_{\mathrm{ML}}(\hat{\mu}_i^{FH})$ can be obtained as

$$\begin{aligned}\widehat{\mathrm{MSE}}_{\mathrm{ML}}(\hat{\mu}_i^{FH}) =& g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2) \\ &- \left(\frac{\sigma_{e,i}^2}{\sigma_v^2+\sigma_{e,i}^2}\right)^2\left(\sum_{j=1}^{m}\frac{1}{(\sigma_v^2+\sigma_{e,j}^2)^2}\right)^{-1} \\ &\times \mathrm{tr}\left(\left(\sum_{j=1}^{m}\frac{\boldsymbol{x}_j\boldsymbol{x}_j^T}{(\sigma_{e,j}^2+\sigma_v^2)}\right)^{-1}\left(\sum_{j=1}^{m}\frac{\boldsymbol{x}_j\boldsymbol{x}_j^T}{(\sigma_{e,j}^2+\sigma_v^2)^2}\right)\right).\end{aligned} \tag{2.61}$$

See Datta and Lahiri (2000), Datta et al. (2005) and Rao and Molina (2015, Chapter 6.2.1) for details.

2.5 The Nested Error Regression Model

While the FH-model presented in the preceding section uses only information on area-level, the second standard SAE model directly models the individual observations. Let y_{ij} be the observation for the jth of n_i elements in area $i, i = 1, \dots, m$ and assume that a data set containing both the variable of interest y_{ij} and a set of unit-level auxiliary information $\boldsymbol{x}_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ is available for the $n = \sum_{i=1}^{m} n_i$ elements in the sample $\mathcal{S}$. Additionally, the p-vector of population means $\bar{\boldsymbol{x}}_{iP} = 1/N_i \sum_{j=1}^{N_i} \boldsymbol{x}_{ij}$ of the covariates is known. Commonly the following model, which is assumed to hold for both the population and the sample, is employed to make inferences on area-level statistics:

$$\begin{aligned} y_{ij} &= \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}, \qquad i = 1, \dots, m, \quad j = 1, \dots, n_i \\ v_i &\overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_v^2) \\ e_{ij} &\overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_e^2). \end{aligned} \tag{2.62}$$

It is assumed that the random effect v_i and the individual error terms e_{ij} are uncorrelated. (2.62) is a simple two-level linear mixed model with an area-specific random intercept v_i. In the context of SAE it is commonly denoted as the (one-fold) nested error regression model or the Battese Harter Fuller (BHF) model after the authors who first applied it to a problem of the discipline (Battese et al., 1988).

The parameters of interest are area-level statistics, e.g. the area means $\overline{Y}_i = 1/N_i \sum_{j=1}^{N_i} y_{ij}$. Under the model specified above, they are given by

$$\overline{Y}_i = 1/N_i \sum_{j=1}^{N_i} (\boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}) = \bar{\boldsymbol{x}}_{iP}^T \boldsymbol{\beta} + v_i + \bar{e}_i, \tag{2.63}$$

with $\bar{e}_i = 1/N_i \sum_{j=1}^{N_i} e_{ij}$. However, in case of large N_i the average error $\bar{e}_i \approx 0$, so that the target statistic is often defined as $\mu_i = E(\overline{Y}_i | v_i) = \bar{\boldsymbol{x}}_{iP}^T \boldsymbol{\beta} + v_i$ (Battese et al. (1988), Pfeffermann (2013), Rao and Molina (2015, Chapter 7.1.1)). Rao and Molina (2015, Chapter 7.1.3) further point out that for small sampling fractions n_i/N_i the EBLUP of $\overline{Y}_i$ approaches the EBLUP of μ_i, thereby giving an additional justification for this simplifying assumption. They also provide an estimator for the case where the sampling fraction is non-negligible.

To deduce details on parameter estimation, prediction and MSE estimation from the results for the general model, the nested error regression model is presented as a special form of the GLMM presented in Section 2.3. It is (see Rao and Molina,

2015):

$$\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i}), \qquad \boldsymbol{X}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}]^T, \qquad \boldsymbol{U}_i = \mathbf{1}_{n_i} \tag{2.64}$$

$$\boldsymbol{v}_i = v_i, \qquad \boldsymbol{\epsilon}_i = (e_{i1}, \ldots, e_{in_i})^T, \qquad \boldsymbol{G} = \sigma_v^2, \qquad \boldsymbol{R}_i = \sigma_e^2 \boldsymbol{I}_{n_i},$$

and thus

$$\boldsymbol{V}_i = \sigma_e^2 \boldsymbol{I}_{n_i} + \sigma_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \qquad \text{and} \qquad \boldsymbol{V}_i^{-1} = \frac{1}{\sigma_e^2} \left(\boldsymbol{I}_{n_i} - \frac{\sigma_v^2}{\sigma_e^2 + n_i \sigma_v^2} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \right).$$

As in the case of the FH-model, the parameter of interest $\mu_i = \bar{\boldsymbol{x}}_{iP}^T \boldsymbol{\beta} + v_i$ is a special case of the general parameter $\eta_i = \boldsymbol{l}_i^T \boldsymbol{\beta} + \boldsymbol{m}_i^T \boldsymbol{v}$ with $\boldsymbol{l}_i = \bar{\boldsymbol{x}}_{iP}^T$ and $\boldsymbol{m}_i = 1$. Making the corresponding substitutions in (2.23) the BLUP for μ can, thus, be derived as

$$\begin{aligned} \tilde{\mu}_i^{\text{BHF}} &= \bar{\boldsymbol{x}}_{iP}^T \tilde{\boldsymbol{\beta}} + \tilde{v}_i^{\text{BLUP}} && (2.65) \\ &= \bar{\boldsymbol{x}}_{iP}^T \tilde{\boldsymbol{\beta}} + \gamma_i (\bar{y}_i - \bar{\boldsymbol{x}}_{iS}^T \tilde{\boldsymbol{\beta}}) && (2.66) \\ &= \hat{\gamma}_i \left(\bar{y}_i + (\bar{\boldsymbol{x}}_{iP} - \bar{\boldsymbol{x}}_{iS})^T \tilde{\boldsymbol{\beta}} \right) + (1 - \gamma_i) \bar{\boldsymbol{x}}_{iP}^T \tilde{\boldsymbol{\beta}}, && (2.67) \end{aligned}$$

where

$$\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \frac{\sigma_e^2}{n_i}}, \tag{2.68}$$

and $\bar{\boldsymbol{x}}_{iS} = 1/n_i \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}$ denotes the p-vector of means of covariates for elements in $\mathcal{S}_i$. Further, (2.9) simplifies to

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i \right), \tag{2.69}$$

with

$$\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i = \frac{1}{\sigma_e^2} \left(\sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T - \gamma_i n_i \bar{\boldsymbol{x}}_{iS} \bar{\boldsymbol{x}}_{iS}^T \right), \tag{2.70}$$

$$\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i = \frac{1}{\sigma_e^2} \left(\sum_{j=1}^{n_i} \boldsymbol{x}_{ij} y_{ij} - \gamma_i n_i \bar{\boldsymbol{x}}_{iS} \bar{y}_{iS} \right), \tag{2.71}$$

and $\bar{y}_{iS} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$.

Equation (2.67) shows that as the FH-estimator, the BHF-estimator can also be interpreted as a composite estimator. It is a linear combination of the survey regression estimator and a synthetic regression estimator. Weights are given by γ_i, i.e. by the share of model variance from the overall variance.

Note, that the BLUP in unsampled areas, i.e. areas with $n_i = 0$, is $\tilde{\mu}_i^{\text{BHF}} = \bar{\boldsymbol{x}}_{iP}^T \tilde{\boldsymbol{\beta}}$. It thus requires $\bar{\boldsymbol{x}}_{iP}^T$ to be available for unsampled areas, too.

The EBLUP $\hat{\mu}_i^{\text{BHF}}$ is obtained from (2.62) by replacing the variance components $\boldsymbol{\vartheta} = (\sigma_v^2, \sigma_e^2)^T$ by an estimate $\hat{\boldsymbol{\vartheta}}$. As in the general case, variance components can be estimated by ML or REML. The asymptotic covariance matrix $\overline{\boldsymbol{V}}$ of $\hat{\boldsymbol{\vartheta}}_{\text{ML}}$ and $\hat{\boldsymbol{\vartheta}}_{\text{REML}}$ is obtained as the inverse of the Fisher information matrix $\boldsymbol{\mathcal{I}}$, which in the case of the nested error regression model is a 2×2-matrix. It is

$$\overline{\boldsymbol{V}}(\hat{\boldsymbol{\vartheta}}^2_{\text{ML}}) = \overline{\boldsymbol{V}}(\hat{\boldsymbol{\vartheta}}^2_{\text{REML}}) = \boldsymbol{\mathcal{I}}^{-1} = \begin{pmatrix} I^{vv} & I^{ve} \\ I^{ev} & I^{ee} \end{pmatrix}, \tag{2.72}$$

with (Datta and Lahiri, 2000)

$$I^{vv} = \frac{2}{a} \sum_{d=1}^{m} \left(\frac{n_d - 1}{\sigma_e^4} + \frac{1}{h_d^2} \right), \tag{2.73}$$

$$I^{ee} = \frac{2}{a} \sum_{d=1}^{m} \frac{n_d^2}{h_d^2}, \tag{2.74}$$

$$I^{ve} = I^{ev} = -\frac{2}{a} \sum_{d=1}^{m} \frac{n_d}{h_d^2}, \tag{2.75}$$

where

$$a = \left(\sum_{d=1}^{m} \frac{n_d^2}{h_d^2} \right) \left(\sum_{d=1}^{m} \left(\frac{n_d - 1}{\sigma_e^4} + \frac{1}{h_d^2} \right) \right) - \left(\sum_{d=1}^{m} \frac{n_d}{h_d^2} \right)^2 \quad \text{and} \quad h_d = \sigma_e^2 + n_d \sigma_v^2.$$

See Rao and Molina (2015, Chapter 7.1.2) for alternative estimation approaches for $\boldsymbol{\vartheta}$ and respective results for the covariance matrix.

The MSE of the BHF estimator, $\text{MSE}(\hat{\mu}_i^{\text{BHF}}) = E(\hat{\mu}_i^{\text{BHF}} - \mu_i)^2$, can be obtained based on general results from 2.3.4. As above (see (2.31)), the MSE of $\hat{\mu}_i^{\text{BHF}}$ is approximated by (Prasad and Rao, 1990)

$$\text{MSE}(\hat{\mu}_i^{\text{BHF}}(\sigma_v^2, \sigma_e^2)) = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2). \tag{2.76}$$

with

$$g_{1i}(\sigma_v^2, \sigma_e^2) = (1 - \gamma_i)\sigma_v^2, \tag{2.77}$$

$$g_{2i}(\sigma_v^2, \sigma_e^2) = (\overline{\boldsymbol{x}}_{iP} - \gamma_i \overline{\boldsymbol{x}}_{iS})^T \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} (\overline{\boldsymbol{x}}_{iP} - \gamma_i \overline{\boldsymbol{x}}_{iS}). \tag{2.78}$$

A second order approximation of g_3 for σ_v^2 estimated by ML or REML was provided by Datta and Lahiri (2000) as

$$g_{3i}(\sigma_v^2, \sigma_e^2) = \frac{1}{n_i^2(\sigma_v^2 + \frac{\sigma_e^2}{n_i})^3}(\sigma_e^4 I^{vv} + \sigma_e^4 I^{ee} - 2\sigma_e^2 \sigma_v^2 I^{ve}). \tag{2.79}$$

Finally, in analogy to the area-level case, a second-order unbiased estimator with $\text{MSE}(\hat{\mu}_i^{BHF})$ for σ_v^2 estimated by REML is given by $\widehat{\text{MSE}}(\hat{\mu}_i^{\text{BHF}}) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$. See Datta and Lahiri (2000) for details and Rao and Molina (2015, Chapter 7.2.2) for results for other estimation methods.

Chapter 3

Finite Mixture Models

3.1 Introduction

Finite Mixture Model (FMM) are a large and rapidly evolving research area. The general framework comprises a broad variety of models, that are applied to statistical problems in diverse disciplines such as biometrics, medicine, genetics, machine learning, marketing, economics and finance. With this wide application in different disciplines, often with specific focusses and diverse naming and notation of central concepts, the field of FMM has evolved to be a vast and often hardly connected research area. See Lindsay (1995) for an overview of different "names" and conceptualizations of the mixture framework.

In their most natural and accessible interpretation, FMM offer an intuitively appealing approach when it is plausible to assume that there is a certain number of – actually existing – subgroups in the population yet subgroup identity is unobserved for all observations (see e.g. Frühwirth-Schnatter (2006, Chapter 1.1), Lindsay (1995, Chapter 1.1), McLachlan and Peel (2000, Chapter 1.4)). The aim of mixture modelling then might either be to appropriately model the distribution of this heterogeneous population or to estimate a model for each subgroup as well as to infer the relative subgroup sizes from the data. In some applications, there might also be an interest in clustering, i.e. in attributing subgroup-identity or a probability of subgroup identity to specific observations. With this intuitive background of *a priori* existing, latent subgroups, the framework of finite mixture modelling is usually conceptualized as a missing data problem, where the realizations of a multinomial variable indicating class membership is missing for all observations (see McLachlan and Peel, 2000, pp. 7, 19–20).

Yet, the assumed components of the mixture distribution do not necessarily cor-

respond with actually existing subgroups. Alternatively, FMM can be interpreted and employed as a flexible semi-parametric way to model unknown or unsmooth distributional shapes without attributing it to an underlying grouping structure (see Frühwirth-Schnatter (2006, pp. 5–6), McLachlan and Peel (2000, pp. 7–8)). Mixtures are apt to capture features possibly present in real data sets, such as heavy tails, skewness or multimodality (see Marron and Wand, 1992, for a presentation of the broad spectrum of shapes a mixture of univariate normal densities can take).

The gain in flexibility, naturally, comes with a price. While mixtures as a modelling framework are an intuitively appealing extension of standard statistical models, they have certain peculiar properties that complicate inference. Wasserman (2012, August 4), therefore, concludes that mixture models are "strange beasts" and jokes that "they should be avoided at all cost". While the peculiarities of the modelling approach should obviously be borne in mind, there are, however, workable and well-established solutions for most of the issues that arise with employing mixture models. The wide use of mixtures in many different areas show that authors are willing to embrace the additional complexity for the advantages of a flexible and intuitive modelling framework.

In the following sections, the framework of finite mixture models and finite mixture regression models is introduced.

3.2 Literature Review

Most notably since the introduction of the EM algorithm by Dempster, Laird and Rubin (1977) (see McLachlan and Krishnan, 2008), finite mixture model theory has received growing attention. See Everitt and Hand (1981), Frühwirth-Schnatter (2006), Lindsay (1995), McLachlan and Basford (1988), McLachlan and Peel (2000), and Titterington, Smith and Makov (1985) for extensive overviews of the theoretical and practical aspects of finite mixture modelling.

With the approach proposed in this thesis, a finite mixture of mixed-effects regression models is considered. Therefore, several works that have previously combined these two modelling approaches can be relied upon in the development of the suggested method. Mainly in the fields of biology and the health sciences, but also in that of marketing, there are a number of applications in which mixtures of mixed-effects models have been utilized (Celeux, Martin and Lavergne, 2005; Lenk and DeSarbo, 2000; Martella, Vermunt, Beekman, Westendorp, Slagboom and Houwing-Duistermaat, 2011; Martinez, Lavergne

and Trottier, 2009; McLachlan, Ng and Wang, 2008; Ng and McLachlan, 2014; Ng, McLachlan, Wang, Ben-Tovim Jones and Ng, 2006; Wang, Ng and McLachlan, 2012; Yau, Lee and Ng, 2003). Verbeke and Lesaffre (1996) and Xu and Hedeker (2001) analysed linear mixed models in which the random effects are distributed according to a mixture of normal distributions (also see Verbeke and Molenberghs, 2000, Chapter 12). Scharl, Grün and Leisch (2010) compared the performance of mixtures of linear regression models, both with and without random effects, in a simulation study. Celeux et al. (2005) and Grün (2008) provide an EM algorithm for the estimation of mixtures of mixed models. In a recent paper, Du, Kahili, Neslehova and Steele (2013) proposed a penalized likelihood approach to model selection for finite mixtures of linear mixed models. Note that, in both these applications and theoretical discussions of finite mixture models, the interest usually lies either in clustering or, less frequently, in the interpretation of component-specific model coefficients, whereas the approach presented in this thesis focuses on predicting a statistic from the estimated model.

Prediction using mixture models has generally received little systematic theoretical consideration. Recently, Cole and Bauer (2016) discussed what they referred to as individual prediction from mixture models. Borrowing concepts from mixed model theory (Skrondal and Rabe-Hesketh, 2009), the authors distinguish between (1) marginal prediction, which averages over unobserved class membership to derive an "overall prediction" (Cole and Bauer, 2016, p. 617), and (2) individual or conditional prediction, which takes individual predictions for latent class membership into account. In an SAE application of mixture models, as proposed in this thesis, these conditional predictions are of interest because mixtures are employed to model areas from actually existing but unobserved groups, each of which has a specific relationship between response variable and covariates. Furthermore, the interest lies in making an area-specific prediction, using all of the information available concerning the area in question. Prediction should thus clearly take the (predicted) area-specific group-membership into account.

Finite mixture models have been extended to include covariates to model the mixture weights (see Dayton and Macready (1988); Farewell (1982); Jacobs, Jordan, Nowlan and Hinton (1991). In addition, see Gormley and Murphy (2011); Ng and McLachlan (2014); Peng, Jacobs and Tanner (1996); Thompson, Smith and Boyle (1998); Wedel (2002); Wedel and Kamakura (2000); Yuksel, Wilson and Gader (2012)). Mainly in the context of market segmentation (Kopsch (2001); Leeflang, Wittink, Wedel and Naert (2000); Wedel (2002); Wedel and Kamakura (2000)), but also in other research areas (Grilli, Rampichini and Varriale, 2015; Grün, 2008; Thompson et al., 1998), the submodels for the mixture weights are referred

to as concomitant variable models, and the approach is referred to as the concomitant variable mixture (regression) model. Mixtures of regression models with covariate-dependent mixture weights are alternatively well-known as mixture-of-experts models (Frühwirth-Schnatter, 2006; Gormley and Murphy, 2011; Peng et al., 1996; Yuksel et al., 2012)), which were originally introduced in the neural network literature by Jacobs et al. (1991). Regardless of its name, this extended approach is particularly useful if the interest lies not only in controlling for heterogeneity as a nuisance in the data but also in identifying and characterizing subgroups in a meaningful way (Leeflang et al., 2000; Wedel and Kamakura, 2000). If suitable covariates are available, this approach also supports the assignment to subgroups. Moreover, the results of the submodel can be used to classify new observations on the basis of the covariates alone (Wedel and Kamakura, 2000). In this thesis, a corresponding extension of finite mixture regression models for SAE is also provided. In addition to the advantages listed above, the improved assignment to subgroups also enhances estimation accuracy. Furthermore, the option of assigning new observations to subgroups based only on the estimated submodel and the covariates, can be employed to predict the statistic of interest for unsampled areas in a heterogeneous population.

3.3 Model Definitions

3.3.1 Finite Mixture Models

Consider a random vector $\boldsymbol{y}$. $\boldsymbol{y}$ is said to arise from a K-component finite mixture distribution if the respective density can be represented by a weighted sum of K component-specific densities $f_k(\boldsymbol{y})$, i.e.

$$f(\boldsymbol{y}) = \sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{y}). \tag{3.1}$$

Here, λ_k are the mixing proportions or mixture weights with $\sum_{k=1}^{K} \lambda_k = 1$ and $\lambda_k > 0$ for all $k = 1, \ldots, K$. Typically (but not necessarily), the K component densities $f_k(\boldsymbol{y})$ are assumed to arise from the same parametric distribution family parametrized by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$, where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ are the K vectors of component-specific model parameters:

$$f(\boldsymbol{y}|\boldsymbol{\psi}) = \sum_{k=1}^{K} \lambda_k f(\boldsymbol{y}|\boldsymbol{\theta}_k). \tag{3.2}$$

$\boldsymbol{\psi} = (\lambda_1, \ldots, \lambda_{K-1}, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$ is a vector containing all the unknown parameters in the mixture distribution. An important and prominent example is a mixture of K multivariate normal distributions $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with component-specific mean $\boldsymbol{\mu}_k$ and covariance-matrix $\boldsymbol{\Sigma}_k$. Then f_k in (3.1) is $f_{\mathcal{N}}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $f_{\mathcal{N}}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the density of the multivariate normal distribution for component k.

As stated previously, the framework of FMM is often conceptualized as a missing-data situation that arises when $\boldsymbol{y}$ is sampled from a population that consists of a certain number of subgroups and subgroup membership is unobserved for all observations (see Frühwirth-Schnatter (2006, Chapter 1), McLachlan and Peel (2000, Chapter 1.4 and 1.9)): Consider a heterogeneous population consisting of K (actually existing) classes and assume that a sample of size n is drawn. Denote the n-dimensional vector of observations by $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. For $i = 1, \ldots, n$ let there, further, be a random subgroup-label vector $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})^T$ indicating class membership of observation i by taking the value 1 if y_i stems from component k and 0 otherwise. $\boldsymbol{z}_i$ follows a multinomial distribution consisting of one draw from the categories $1, \ldots, K$ with probabilities $\lambda_1, \ldots, \lambda_K$. Note that in this framework the mixing proportions λ_k can be interpreted as the relative subgroup sizes. They alternatively might be understood as the unconditional probability that an observation belongs to class k, that is $\lambda_k = Pr(z_{ik} = 1)$ for all i. For completeness, define $\boldsymbol{z} = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T)^T$.

Assume that the conditional density of $\boldsymbol{y}$ given $z_{ik} = 1$ can be represented by the component-specific density $f_k(\boldsymbol{y})$. Then, the marginal density of $\boldsymbol{y}$ is given by the mixture density as specified in (3.1). The mixture framework, thus, naturally arises if the component labels are unobserved, i.e. $\boldsymbol{z}_i$ is missing for all observations in the sample, and only $\boldsymbol{y}$ is recorded.

3.3.2 Finite Mixtures of Regression Models

An obvious extension of FMM are Finite Mixtures of Linear Regression Models. Standard regression models assume a constant linear relationship between response variable and covariates. This assumption might, however, be overly restrictive. If so, Finite Mixtures of Linear Regression Models, that allow for different sets of regression parameters between a fixed number of different unobserved (or unspecified) subgroups of observations in statistical modelling, might be a suitable framework.

Early examples of this model class can be found in biology (Hosmer, 1974), marketing (DeSarbo and Cron, 1988) and economics (Fair and Jaffee, 1972; Quandt, 1972; Quandt and Ramsey, 1978). Due to their flexibility, Finite Mixtures of Regression Models, nowadays, are in wide use in a broad range of

disciplines. Depending on the specific research area, they are also referred to as switching regression models, finite regression mixtures and latent class regression models. See Frühwirth-Schnatter (2006, Chapter 8), Grün and Leisch (2008) and Wedel and Desarbo (2002) for reviews. R-packages for fitting and analyzing mixtures of regression models are provided with `flexmix` (Grün and Leisch, 2007; Leisch, 2004) and `mixtools` (Benaglia, Chauveau, Hunter and Young, 2009).

Assume that the observed values $\boldsymbol{y} = (y_i, \ldots, y_n)^T$ depend on a set of covariates $\boldsymbol{X}$ in a linear way

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{3.3}$$

to obtain a standard linear regression model. Now loose the assumption of fixed regression coefficients for all observations: Let there be K different sets of model parameters, $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T, \boldsymbol{\vartheta}_k^T)^T$, each valid in a subgroup of the population. For each observation i, let there further be a latent random subgroup-label vector $\boldsymbol{z_i}$, that determines subgroup membership of i and, thus, the set of model parameters. Under this scenario, the conditional density of $\boldsymbol{y}$ given $\boldsymbol{X}$ is a mixture of the K component densities $f_{\mathcal{N}}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k)$:

$$f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\psi}) = \sum_{k=1}^{K} \lambda_k f_{\mathcal{N}}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k) \tag{3.4}$$

Note that (3.4) is a special form of the mixture of multivariate gaussians with covariate-dependent mean vector $\boldsymbol{\mu}_k = \boldsymbol{X}\boldsymbol{\beta}_k$.

3.3.3 Modelling the Mixture Weights

Finite Mixture Models have been extended to include covariates for the mixture weights, λ_k. See Dayton and Macready (1988), Farewell (1982), Jacobs et al. (1991) for early suggestions under different names, Ng and McLachlan (2014); Peng et al. (1996); Thompson et al. (1998) for examples of applications and Gormley and Murphy (2011), McLachlan and Peel (2000, Chapter 5.5.1), Wedel and Kamakura (2000), Wedel (2002), Yuksel et al. (2012)) for reviews. This supports partitioning the observations into subgroups when the span of the covariates is different between the components. Further, it helps to characterize and analyse the subgroups in a meaningful way thereby providing insights into the structures present in the data (Leeflang et al., 2000; Wedel and Kamakura, 2000). Finally, the submodel allows for an out-of-sample prediction of subgroup membership for new observations on basis of the covariates only (Wedel and Kamakura, 2000).

The covariates for the mixture weights, which might or might not be distinct to the covariates in the component model, are frequently denoted as concomitant variables (Grilli et al., 2015; Grün, 2008; Thompson et al., 1998; Wedel, 2002). Correspondingly, the approach is commonly referred to as concomitant variable mixture model or concomitant variable mixture regression models, if the component densities are covariate-dependent, too. In this latter form, the framework is also well-known as mixtures-of-experts model (Frühwirth-Schnatter (2006); Gormley and Murphy (2011); McLachlan and Peel (2000); Peng et al. (1996); Yuksel et al. (2012, Chapter 5.13.1)). This model from the neural network literature dates back to Jacobs et al. (1991). In the respective literature, the component models are referred to as experts and the models for the mixing proportions are known as gating networks. See Gormley and Murphy (2011) or Yuksel et al. (2012) for a review.

In what follows a finite mixture regression model with concomitant variables is considered, i.e. the mixture regression model as defined in (3.4) is extended to include covariates for the mixture weights:

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\psi}) = \sum_{k=1}^{K} \lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha}) f_{\mathcal{N}}(y; \boldsymbol{x}_i, \boldsymbol{\theta}_k), \quad i = 1, \ldots, n. \tag{3.5}$$

$\lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha}) = \lambda_{i,k}$ with $\sum_{k=1}^{K} \lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha}) = 1$ and $\lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha}) > 0$ for all k denotes the mixture weight or the prior probability that an observation i belongs to component k. These weights are assumed to be functionally related to a set of auxiliary information $\boldsymbol{w}_i$ through a submodel, typically a multinomial logit model

$$\lambda_{i,k} = \lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{w}_i^T \boldsymbol{\alpha}_k)}{\sum_{j=1}^{K} \exp(\boldsymbol{w}_i^T \boldsymbol{\alpha}_j)}, \tag{3.6}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{K-1}^T)^T$ and $\boldsymbol{\alpha}_K = \mathbf{0}$ for identifiability. Thus, $\boldsymbol{\psi}$, i.e. the vector containing all parameters in the mixture distribution, now contains both the parameters of the sub- and the main model: $\boldsymbol{\psi} = (\boldsymbol{\alpha}^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)$.

Note that the extended mixture regression model with submodel for the mixture weights contains the mixture regression model as a special case: When $\boldsymbol{w}_i = 1$ for all i, $\lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha}_k) = \lambda_k$ is just a component-specific weight as in the standard mixture regression model defined in (3.4)

As detailed above, the motivation for estimating a mixture model with concomitant variables might be to further analyse and describe the derived partition of the data into clusters. In this case, instead of the integrated approach of simultaneously estimating the mixture components and modelling the mixture weights as functions of concomitant variables a stepwise procedure can be applied to characterize the

subgroups. After fitting a standard finite mixture regression model, either the estimated posterior probabilities or predicted subgroup assignments based on these posterior probabilities can be regressed on the covariates in a subsequent step (Gudicha and Vermunt, 2013; Wedel, 2002). This approach does not account for neither the estimation error of the posterior probabilities nor the classification error, if the observations are clustered using some clustering rule (Gudicha and Vermunt, 2013; Leeflang et al., 2000). Furthermore, the submodel for the component membership is estimated independently of the component densities, optimizing the sum of squared errors in $\xi_{i,k}$ instead of the mixture likelihood. As Leeflang et al. (2000) point out, the assignment derived, therefore, "[does] not possess an 'optimal' structure with respect to [its] profile on the concomitant variables". Consequently, the effects of the covariates in the submodel are severely underestimated (see Gudicha and Vermunt, 2013, for simulation results). Note, however, that the estimation of parameters in the submodel is not necessarily of first priority in the intended application of mixture models in SAE. The assessment of the competing methods in this context also has to include their performance with respect to assigning the observations into subgroups and, most importantly, predicting the variable of interest.

3.4 Identifiability

Meaningful inference on a model's parameters is only possible if the model is identifiable. Generally, a parametric family of distributions, represented by the corresponding family of densities $\{f(\boldsymbol{y}|\boldsymbol{\psi}) : \boldsymbol{\psi} \in \boldsymbol{\Omega}\}$, where $\boldsymbol{\Omega}$ denotes the parameter space, is said to be identifiable if distinct values of the parameter determine distinct members of the family. Put differently, for two parameters $\boldsymbol{\psi}$ and $\boldsymbol{\psi}^*$, $f(\boldsymbol{y}|\boldsymbol{\psi})$ and $f(\boldsymbol{y}|\boldsymbol{\psi}^*)$ are identical for almost every $\boldsymbol{y}$, if and only if $\boldsymbol{\psi} = \boldsymbol{\psi}^*$. Thus, if an infinite number of realizations from the model could be observed, it would be possible to learn the true (and unique) values of the parameters.

In the case of finite mixture models, the question of identifiability is somehow more complex. Three different types of identifiability problems can be distinguished:

A first type of nonidentifiability is due to a possible rearrangement or relabeling of components. As noted by Redner and Walker (1984) the mixture density is invariant to the change of component labels in $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ and $(\lambda_1, \ldots, \lambda_K)$ if the component densities belong to the same parametric family. In fact, for a mixture of K components from the same parametric family there are $K!$ different ways of arranging the components, each giving rise to the same mixture distribution (see Frühwirth-Schnatter (2006, Chapter 1.3) and McLachlan and Peel

(2000) for further details). Therefore, finite mixture models are not identifiable in the strict sense of the general definition given above and commonly an adapted, weaker definition of identifiability is applied (see e.g. Frühwirth-Schnatter, 2006; McLachlan and Peel, 2000; Yakowitz and Spragins, 1968): A FMM is said to be identifiable if for any two vectors of parameters $\boldsymbol{\psi}, \boldsymbol{\psi}^* \in \boldsymbol{\Omega}$ the equality

$$\sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{y}|\boldsymbol{\theta}_k) = \sum_{k=1}^{K^*} \lambda_k^* f_k(\boldsymbol{y}|\boldsymbol{\theta}_k^*) \tag{3.7}$$

for almost every $\boldsymbol{y}$, implies that $K = K^*$ and that $\boldsymbol{\psi}^*$ can be permuted such that $\boldsymbol{\psi} = \boldsymbol{\psi}^*$.

Note that nonidentifiability due to the rearrangement of component labels can be considered as an inconvenience that has to be borne in mind by the researcher, but does not really pose a problem in standard applications as long as a frequentist estimation approach is taken. It might however cause difficulties in a Bayesian framework when inferences are obtained using posterior simulations. The same is true for simulation studies carried out to evaluate the performance of competing statistical methods as in Chapter 5. The issue is sometimes solved by imposing an adequate constraint on the solution such as requiring $\lambda_1 < \lambda_2 < \ldots < \lambda_k$ (Aitkin and Rubin, 1985) or by imposing an order constraint on a selected element of the component parameter vectors $\boldsymbol{\theta}_k$. Obviously, such a restriction has to be chosen carefully in order to fulfil its purpose in all possible constellations (see Frühwirth-Schnatter, 2006, Chapter 1.33 for a critical discussion). In the simulations studies carried out in the course of this work, the issue was solved by ordering the components according to the size of the estimated intercepts.

A second source of lack of identifiability is commonly discussed as nonidentifiability due to overfitting. As noted by Crawford (1994), any finite mixture model with K components can also be represented by a mixture of $K + 1$ components, where either one component is empty (i.e. $\lambda_k = 0$) or two components have the identical set of parameters (see Frühwirth-Schnatter, 2006, Chapter 1.3.2 for details). This kind of unidentifiability problem can, however, be easily solved by imposing adequate constraints on the parameters: It is usually required that the mixture weights are larger than zero and that the K sets of component parameters are distinct in the weak sense that any two parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ differ in at least one element (see Frühwirth-Schnatter, 2006, Chapter 1.3.3).

Finally, there is a third type of identifiability problem, i.e. the lack of what is sometimes denoted as "generic identifiability" (Frühwirth-Schnatter, 2006, p. 22): Finite mixture models might still be unidentifiable, even after imposing adequate constraints on the mixture weights and under the weak definition given above. For example, as first pointed out by Teicher (1961), mixtures of

uniform distributions are generally unidentifiable. Some literature is devoted to identifiability or conditions of identifiability for specific families of FMM (Ahmad, 1988; Feller, 1943; Holzmann, Munk and Gneiting, 2006; Teicher, 1961,6; Yakowitz and Spragins, 1968). The identifiability of mixtures of multivariate normal distributions was proved by Yakowitz and Spragins (1968). Hennig (2000) showed that the identifiability of mixtures of linear regression models with normally distributed errors does not readily follow from this result. He did, however, demonstrate that mixtures of linear regression models are identifiable under sufficient conditions easily fulfilled in many applications. More specifically, a mixture of linear regression models is identifiable if the number of components is smaller than the minimum number of $(h_x - 1)$-dimensional hyperplanes formed by the design points, with h_x denoting the number of predictors in $\boldsymbol{X}$ excluding the intercept. Based on this result, Du et al. (2013) accordingly formulate the sufficient condition for the identifiability of mixtures of regression models with mixed effects: A mixture of LMMs with design matrices $\boldsymbol{X}$ and $\boldsymbol{U}$ is identifiable if the number of mixture components is smaller than the number of $(h_Q - 1)$-dimensional hyperplanes required to cover the design points in $\boldsymbol{Q}$, where $\boldsymbol{Q}$ is the matrix formed by the distinct columns of the matrix $[\boldsymbol{X}, \boldsymbol{U}]$ and h_Q is the number of columns in $\boldsymbol{Q}$. Thus, identifiability problems might arise if the variability of the design points is low. As Hennig (2000) points out, the identifiability condition is usually fulfilled in practical applications. He further argues, that it might be problematic if the covariates can take only a limited number of values, i.e. if dummy variables or answers from a questionnaire with a small number of possible answers are considered as covariates. Finally, Jiang and Tanner (1999) study the identifiability of mixture models with submodel for the mixture weights. They show that models in this class are identifiable under certain regularity conditions, provided $\boldsymbol{\alpha}_K = \boldsymbol{0}$ is assumed. They further prove that the conditions are fulfilled for mixtures with univariate normal components.

In all theoretical elaborations that follow, it is assumed that the mixture model under consideration is identifiable in the weaker sense commonly applied in the context of FMM.

3.5 Parameter Estimation

Parameter estimation is performed using the the frequentist approach of ML estimation. Alternatively, a Bayesian approach can be adopted. See McLachlan and Peel (2000, Chapter 4) or Frühwirth-Schnatter (2006, Chapter 3 and 5) for an overview of Bayesian inference for FMM.

The likelihood function for $\boldsymbol{\psi}$ under a finite mixture model is obtained from the joint density of $\boldsymbol{y}$ as

$$L(\boldsymbol{\psi}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \lambda_k f(y_i|\boldsymbol{\theta}_k). \tag{3.8}$$

The corresponding log-likelihood $l(\boldsymbol{\psi})$ is given by

$$l(\boldsymbol{\psi}) = \log(L(\boldsymbol{\psi})) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \lambda_k f(y_i|\boldsymbol{\theta}_k) \right). \tag{3.9}$$

Maximum likelihood estimates for the parameters $\boldsymbol{\theta}_k$ as well as the model probabilities λ_k are obtained by maximizing this log-likelihood given the observed realizations for $\boldsymbol{y}$. However, optimization of (3.9) can not be done directly because the log of a sum in the function makes its derivative computationally intractable. Maximum likelihood estimates are, therefore, obtained using the EM-algorithm, a general-purpose optimization algorithm generally introduced in Appendix A.1. See McLachlan and Peel (2000) and McLachlan and Krishnan (2008) for an extensive overview of parameter estimation via the EM algorithm for FMM.

In this estimation context, the FMM framework is interpreted as a missing data problem. As described in Section 3.3.1, this implies the notion that each observation y_i belongs to one of the K classes with $\boldsymbol{z}_i$ indicating the true class membership for observation y_i. The complete data set would, thus, contain n realizations of y_i and $\boldsymbol{z}_i$ and the complete-data log-likelihood l_c is given by

$$l_c(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} z_{ik}\, \lambda_k f(y_i|\boldsymbol{\theta}_k) \right), \tag{3.10}$$

which can be rearranged[1] as

$$l_c(\boldsymbol{\psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} (\log \lambda_k + \log f\,(y_i|\boldsymbol{\theta}_k)). \tag{3.11}$$

Maximizing l_c with respect to the model parameters $\boldsymbol{\theta}_k$, the model probabilities λ_k as well as the latent variable $\boldsymbol{z}$ is performed iteratively by altering between the E-step, where a conditional expectation Q of $l_c(\boldsymbol{\psi})$, given $\boldsymbol{y}$ and parametrized

[1] For any i, $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$ takes the value 0 in $K-1$ cases and the inner sum $\sum_{k=1}^{K} z_{ik}\,(\lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha}_k) f\,(y_i|\boldsymbol{\theta}_k))$ reduces to a single term $\lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha}_k) f(y_i|\boldsymbol{\theta}_k)$ with corresponding log given by $\log \lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha}_k) + \log f\,(y_i|\boldsymbol{\theta}_k))$.

by the current estimates of the unknown parameters in $\boldsymbol{\psi}$ is derived and the M-step where an updated estimate for $\boldsymbol{\psi}$ is obtained by maximizing this conditional expectation of l_c.

More detailed, in the context of FMM the following algorithm is applied:

- *Specification of starting values*
  To begin, an initial choice of starting values is made. This can either be an initial assumption for the parameters $\boldsymbol{\psi}$ or a partition of observations into K groups. See McLachlan and Peel (2000, Chapter 2.12) for a discussion of different initialization strategies and related convergence properties.

- *E-step*
  The conditional expectation of the complete-data log-likelihood $l_c(\boldsymbol{\psi})$ given $\boldsymbol{y}$ under the current estimate for $\boldsymbol{\psi}$ is derived as

$$\begin{aligned} Q(\boldsymbol{\psi}; \widehat{\boldsymbol{\psi}}^{(t-1)}) &= E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}[l_c(\boldsymbol{\psi})|\boldsymbol{y}] \\ &= E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}(\log\lambda_k + \log f_k\left(y_i|\boldsymbol{\theta}_k\right))\middle|\boldsymbol{y}\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{\xi}_{i,k}^{(t-1)}(\log\lambda_k + \log f_k\left(y_i|\boldsymbol{\theta}_k\right)), \end{aligned} \tag{3.12}$$

  where $\hat{\xi}_{i,k}^{(t-1)}$ denotes the conditional expectation for $z_{ik} = 1$ given y_i and the current estimate for $\boldsymbol{\psi}$. They are calculated using $\widehat{\boldsymbol{\psi}}^{(t-1)}$, i.e. the estimates for $\boldsymbol{\theta}_k$ and λ_k obtained in the last iteration step $(t-1)$ (or, in the first iteration, the chosen starting values):

$$\begin{aligned} \hat{\xi}_{i,k}^{(t-1)} &= Pr_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(z_{ik} = 1|y_i) \\ &= \frac{\hat{\lambda}_k^{(t-1)} f(y_i|\hat{\boldsymbol{\theta}}_k^{(t-1)})}{\sum_{k'\in K}\hat{\lambda}_{k'}^{(t-1)} f(y_i|\hat{\boldsymbol{\theta}}_{k'}^{(t-1)})} \end{aligned} \tag{3.13}$$

- *M-step*
  In the M-step an updated estimate $\widehat{\boldsymbol{\psi}}^{(t)}$ for $\boldsymbol{\psi}$ is obtained by maximizing Q as derived in the E-step. Estimation of $\boldsymbol{\theta}_k$ can be done for each component model separately by maximizing the weighted component-specific

log-likelihood $\sum_{i=1}^{m} \hat{\xi}_{i,k}^{(t)} \log f\left(y_i|\boldsymbol{\theta}_k\right)$. Deriving an update for λ_k requires maximizing $\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{\xi}_{i,k}^{(t)} \log \lambda_k$ under the constraint $\sum_{k=1}^{K} \lambda_k = 1$. The result is

$$\hat{\lambda}_k^{(t)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_{i,k}^{(t-1)}, \tag{3.14}$$

i.e. $\hat{\lambda}_k^{(t)}$ is obtained by taking the average of $\hat{\xi}_{i,k}^{(t-1)}$ over all observations.

- *Termination*
  Both steps are repeated until convergence (see Appendix A.1)

The EM algorithm can also be applied if a submodel for the mixture weights as introduced in Section 3.3.3 is assumed. In this case, λ_k in l_c is replaced by $\lambda_{i,k} = \lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha})$. The expectation of the complete data log-likelihood is then given by

$$Q(\boldsymbol{\psi}; \widehat{\boldsymbol{\psi}}^{(t-1)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{\xi}_{i,k}^{(t-1)} (\log \lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha}_k) + \log f\left(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_k\right)), \tag{3.15}$$

where $\boldsymbol{\psi}$ now contains both the parameters of the sub- and the main model, i.e. $\boldsymbol{\psi} = (\boldsymbol{\alpha}^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)$. $\hat{\xi}_{i,k}^{(t-1)}$, the conditional expectation of class membership is derived using the current fit of individual mixture weights $\hat{\lambda}_{i,k} = \lambda_k(\boldsymbol{w}_i, \widehat{\boldsymbol{\alpha}})$ instead of $\hat{\lambda}_k$, i.e.

$$\hat{\xi}_{i,k}^{(t-1)} = \frac{\hat{\lambda}_k(\boldsymbol{w}_i, \widehat{\boldsymbol{\alpha}}_k^{(t-1)}) f(y_i|\hat{\boldsymbol{\theta}}_k^{(t-1)})}{\sum_{k' \in K} \hat{\lambda}_{k'}(\boldsymbol{w}_i, \widehat{\boldsymbol{\alpha}}_{k'}^{(t-1)}) f(y_i|\hat{\boldsymbol{\theta}}_{k'}^{(t-1)})} \tag{3.16}$$

Updating the current parameter fit in the *M-step* also requires obtaining an estimate for $\boldsymbol{\alpha}$. It is obvious from (3.15) that optimization for the submodel and the main model can be done separately.

If a multinomial logit submodel as defined in (3.6) is assumed to hold for the mixture weights, its parameters, $\boldsymbol{\alpha}$, are estimated employing maximum likelihood estimation of generalized linear model, using $\hat{\xi}_{i,k}^{(t)}$ as response vector ((Dang and McNicholas, 2015; Grün, 2008; McLachlan and Peel, 2000; Thompson et al., 1998; Wedel and Kamakura, 2000). $\hat{\lambda}_{i,k}^{(t)}$ is then predicted from the fitted model. For all calculations in this thesis `multinom` from the R-package `nnet` for parameter estimation (see Venables and Ripley, 2002) was employed for this purpose.

As stated in Appendix A.1, in case of multimodal distributions the EM algorithm might converge to a local maximum depending on the chosen starting values. It is therefore often performed repeatedly with different starting values. In both the simulation study and the application presented below this simple strategy was adopted.

3.6 Estimating the Number of Components

A crucial and much discussed question in estimating finite mixture models is the choice of K. In fact, standard estimation techniques for mixtures are based on the assumption of a known number of components. There are practical applications where the number of subgroups in the population is known *a priori*, but usually K has to be selected based on the available data. Note that this particular question of model selection includes the essential decision between $K = 1$ or $K > 1$, i.e. the question whether it is appropriate to assume latent heterogeneity and to accept the larger complexity of employing a mixture model in the first place.

Before a selection of measures for the choice of K is discussed in detail, note that the examination of the histogram of the sampled data, a simple informal method that might be inspired by striking plots in introductory chapters of monographs on mixtures (see for example the famous fishery data example given in Titterington et al. (1985, p. 10)) and its representation in reviews on mixture models (e.g. Frühwirth-Schnatter, 2006, p. 2), might be misleading (Everitt and Hand, 1981, p. 208). Not only can this lead to erroneously assuming a mixture distribution in case of spurious subgroups as McLachlan and Peel (2000, Chapter 1.8) elaborate based on Day (1969). The number of modes in the mixture distribution also does not necessarily correspond to the number of components and in fact there are many examples where it does not. Some families of finite mixture distributions, e.g. mixtures of exponentials, are always unimodal and the number of modes in mixtures of normal distributions depends on the mixture weights and the distance between the component distributions (see McLachlan and Peel (2000, Chapter 1.5) for some instructive examples and Frühwirth-Schnatter (2006, Chapter 1.2.2) or Titterington et al. (1985, Chapter 5.6) for an account of conditions for the number of modes in mixtures of Gaussians).

As Frühwirth-Schnatter (2006, p. 99) points out, estimating K is a difficult problem. The procedure commonly applied is to calculate a suitable criterion, letting K grow. This procedure, obviously, requires inference for an overfitting model, i.e. for a model with more than the true number of components. In this case, the parameters are not identifiable, i.e. central regularity conditions at the heart of

statistical theory underlying frequentist inference are not fulfilled. Thus, the theoretical foundation of model selection techniques commonly used in other contexts do not apply in this particular case (see Frühwirth-Schnatter (2006, Chapter 4.2) or McLachlan and Peel (2000, Chapter 6.4) for details). Selecting K is, however, deemed a central question of estimating finite mixture models so that, despite of these theoretical difficulties, much attention is paid to the task of finding satisfying solutions. As a result, there is a broad variety of different approaches to the question of selecting K and a wide range of measures has been suggested in the literature. See Fonseca and Cardoso (2007); Frühwirth-Schnatter (2006); McLachlan and Peel (2000) and McLachlan and Rathnayake (2014) for reviews. Simulation studies comparing different measures, which due to the lack of thorough theoretical justification of criteria in common use play a vital role in this research field, are given in Biernacki, Celeux and Govaert (1998); Fonseca and Cardoso (2007); Hawkins, Allen and Stromberg (2001); Hettmansperger and Thomas (2000); McLachlan and Ng (2000); Roeder and Wasserman (1997); Sarstedt and Schwaiger (2008) and Windham and Cutler (1992).

Biernacki et al. (1998), Frühwirth-Schnatter (2006, Chapter 7.1.4), McLachlan and Peel (2000, Chapter 6.1) and McLachlan and Rathnayake (2014) argue that the choice of an adequate criterion depends on the purpose of employing a finite mixture model: As stated in the introductory section of this chapter, mixtures can either be seen as a flexible, semiparametric approach to model unknown distributional shapes or as a suitable framework for model-based clustering. In the first case, standard criteria as the Bayesian Information Criterion (BIC) are commonly deemed adequate. If mixture models are applied for model-based clustering, the BIC however tends to overestimate the number of subgroups. As Frühwirth-Schnatter (2006) writes, in this context, a measure is more adequate that takes into account that the "mixture model is fitted with the hope of finding a good partition of the data" (Frühwirth-Schnatter, 2006, p. 213). Thus, in this case a criterion which penalizes poorly separated subgroups, is more suitable.

The objective of the suggested application of mixture models in SAE lies somewhere in between these two purposes: The approach is motivated by the assumed existence of different subgroups of areas. One aim definitely is identifying and describing these groups and a meaningful result in this regard intuitively supports the employment of this more complex modelling approach instead of a standard SAE model. First and foremost, the interest, however, lies in finding a model that matches the data in the case of a heterogeneous population in order to make valid inferences on the statistic of interest. A sharp separation of areas into meaningful subgroups is not necessarily a precondition for this.

Based on these deliberations, suitable measures for each of the two purposes is chosen for further consideration in the present context of selecting the number of components in an application of mixtures of Gaussians in SAE. When conflicting results are obtained, an informed decision considering the larger picture of the specific application has to be made. Given the unresolved issues in the research area, a similar strategy of evaluating and balancing suggestions from different criteria in light of the data situation at hand has for example been recommended by Bauer and Curran (2004) and Nagin (2005).

A common approach to selecting K for the first purpose of density estimation is the consideration of different information criteria (e.g. see Burnham and Anderson, 2002, for a general review of the information-theoretical approach to model selection). See McLachlan and Peel (2000, Chapter 6.8) and McLachlan and Rathnayake (2014) for an overview and discussion of the broad range of respective measures suggested for selecting the number of components. A specific criterion commonly applied is the well-known BIC. It was first suggested by Schwarz (1978) as an approximation to the log integrated likelihood and has evolved to be a standard criterion for model selection.

In the present context of choosing the number of components it can be expressed as

$$\mathrm{BIC} = -2 \log L(\widehat{\boldsymbol{\psi}}) + d \log n, \tag{3.17}$$

where $\widehat{\boldsymbol{\psi}}$ denotes the estimated parameter vector and d is the dimension of $\widehat{\boldsymbol{\psi}}$, i.e. the number of parameters in the model. The first term accounts for the goodness-of-fit of the model under consideration, while the second term is a penalty for model complexity, which increases as the sample size increases. K is chosen to minimize (3.17). A sample-sized adjusted version of the BIC (Sample-Size Adjusted Bayesian Information Criterion (BICadj)) has been proposed by Sclove (1987). It replaces $\log n$ in the second term of (3.17) by $\log\left((n+2)/24\right)$, thereby reducing the sample size penalty.

The theoretical derivation of the BIC as an approximation to the log integrated likelihood relies on identifiability of the model parameters and, therewith, on regularity conditions that are not fulfilled in the present context (see McLachlan and Peel, 2000, Chapter 6.9). Leroux (1992) showed that under mild conditions the BIC does not underestimate K asymptotically. Extending this work and particularly focussing on a potential overestimation of K, Keribin (2000) proved the consistency of the BIC as an estimator of the true number of components under stronger regularity conditions. Moreover, the BIC has proved to perform well in simulation studies (Biernacki et al., 1998; Dasgupta and Raftery, 1998; Roeder and Wasserman, 1997) and there is some support for its usage

in this context (Fraley and Raftery, 2002). Good results for the sample-size adjusted BIC were obtained by Yang and Yang (2007). It further proved to be the most promising selection criterion in a simulation study on choosing K for mixtures of regression models (Sarstedt and Schwaiger, 2008). Both the BIC and its sample-size adjusted version are considered as criteria for selecting K in the context of this thesis.

Secondly, a much considered criterion for choosing K in the context of mixture-based clustering was proposed by Biernacki et al. (1998) based on work of Biernacki and Govaert (1997): They derive a measure based on the integrated complete-data likelihood (also classification likelihood in this context) instead of the likelihood based on the observed data. The resulting criterion, the so-called Integrated Classification Likelihood (ICL), takes the separation of components into account and overcomes the theoretical shortcomings of the BIC. See McLachlan and Peel (2000, Chapter 6.10) and Biernacki et al. (1998) for a more detailed discussion. Let $\tilde{\boldsymbol{z}}$ be an estimator of the true component-label vector $\boldsymbol{z}$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1^T, \ldots, \tilde{\boldsymbol{\theta}}_K^T)^T$ denote the complete-data estimator maximizing $\log p(\boldsymbol{y}|\tilde{\boldsymbol{z}}, \boldsymbol{\theta})$. The ICL for a mixture with K components is given by:

$$\text{ICL} = -2 \log p(\boldsymbol{y}|\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\theta}}) + d_c \log n - 2 \log \frac{\Gamma(K/2) \prod_{k=1}^{K} \Gamma(\tilde{n}_k + \frac{1}{2})}{\Gamma(n + \frac{K}{2}) \Gamma(\frac{1}{2})^K}, \tag{3.18}$$

where d_c denotes the number of distinct elements in $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$ and $\tilde{n}_k$ is an estimate of $n_k = \sum_{i=1}^{n} z_{ik}$, i.e. the number of elements in $k, k = 1, \ldots, K$, based on $\tilde{\boldsymbol{z}}$. Γ denotes the gamma function.

McLachlan and Peel (2000, Chapter 6.10) provide an expression of the ICL for the special case of $\tilde{\boldsymbol{z}} = \hat{\boldsymbol{\xi}}$, where $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\xi}}_1^T, \ldots, \hat{\boldsymbol{\xi}}_n^T)^T$ and $\hat{\boldsymbol{\xi}}_i = (\hat{\xi}_{i,1}, \ldots, \hat{\xi}_{i,K})^T$. Thus, they replace the (estimated) hard component-labels by the respective ML estimates of the conditional expectations of component membership. In this case, maximizing $\log p(\boldsymbol{y}|\tilde{\boldsymbol{z}}, \boldsymbol{\theta})$ yields the ML estimator $\hat{\boldsymbol{\theta}}$, such that $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$, and the ICL can be formulated as

$$\begin{aligned} \text{ICL} = &- 2 \log L(\hat{\boldsymbol{\psi}}) + 2 \,\text{EN}(\hat{\boldsymbol{\xi}}) + 2n \sum_{k=1}^{K} \hat{\lambda}_k \log \hat{\lambda}_k + d_c \log n \\ &- 2 \log \frac{\Gamma(K/2) \prod_{k=1}^{K} \Gamma(n\hat{\lambda}_k + \frac{1}{2})}{\Gamma(n + \frac{K}{2}) \Gamma(\frac{1}{2})^K} \end{aligned} \tag{3.19}$$

Note that now $\tilde{n}_k = n\hat{\lambda}_k$.

Biernacki et al. (1998) derive an approximation to the ICL in case of large enough n_k, showing that the ICL reduces to an "*à la BIC* approximation" (Biernacki

et al., 1998, p. 10) of the logarithm of the complete data integrated likelihood. McLachlan and Peel (2000, Chapter 6.10) provide the respective expression for their ML-version of the ICL (for the general approximation for $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{\theta}}$ see Biernacki et al. (1998, p. 10)) and denote it by ICL-BIC:

$$\begin{aligned} \text{ICL} \approx \text{ICL-BIC} &= -2 \log L(\hat{\boldsymbol{\psi}}) + d \log n + 2\, \text{EN}(\hat{\boldsymbol{\xi}}) \\ &= \text{BIC} + 2\, \text{EN}(\hat{\boldsymbol{\xi}}), \end{aligned} \tag{3.20}$$

where

$$\text{EN}(\boldsymbol{\xi}) = -\sum_{k=1}^{K} \sum_{i=1}^{n} \xi_{i,k} \log \xi_{i,k} \tag{3.21}$$

denotes the entropy of the $n \times K$ fuzzy classification matrix $\boldsymbol{C} = (\xi_{i,k})$ and $\text{EN}(\hat{\boldsymbol{\xi}})$ denotes the respective estimate based on $\hat{\boldsymbol{\xi}}$. It is a measure for the goodness of the partition of the data into clusters: If subgroups are well separated, $\text{EN}(\hat{\boldsymbol{\xi}})$ is close to zero while it takes large values for poorly separated clusters (see Celeux and Soromenho, 1996, who considered the normalized entropy as a criterion for selecting K).

The ICL and its approximation, additional to model complexity, penalize poorly separated components. They, thus, favour a choice of K that leads to a partition of the data into well-separated clusters. These criteria are, therefore, particularly suitable in a clustering context. In the context of this thesis, the ICL-BIC, as defined in (3.20), i.e. the asymptotic version of the measure for ML estimates is considered. Favourable simulation results for this version can be found in McLachlan and Ng (2000). Further, in a simulation study performed by Biernacki et al. (1998), the approximation for large n_k yielded no different results than the ICL in its more accurate form.

3.7 Clustering via Finite Mixture Models

In many applications of FMM there is an interest in classifying the set of n observations y_i into K subgroups. Like in the missing data interpretation of mixtures given in Section 3.3.1, in such a clustering framework, the components of the mixture are assumed to correspond to K subgroups. Observations from subgroup k are distributed according to a component-specific distribution, characterized by the component-density f_k. In comparison to heuristic clustering approaches in common use, model-based clustering employing mixture distributions has the advantage of making underlying (model-)assumptions explicit and allowing for inference

within the framework of well-established statistical theory (see McLachlan and Peel (2000, Chapter 1.15) or Fraley and Raftery (2002) for a short discussion of advantages of model-based clustering and further references). Accordingly, there is a large number of applications of FMM to clustering from different research areas. See Fraley and Raftery (2002), Frühwirth-Schnatter (2006, Chapter 7.1), McLachlan and Basford (1988) and McLachlan and Peel (2000, Chapter 1.15) for reviews and detailed discussions of clustering via FMM.

From Bayes' rule, the conditional or posterior probability that observation i belongs to component k is given by

$$\xi_{i,k} = Pr(z_{ik} = 1|y_i, \boldsymbol{\psi}) = \frac{\lambda_k f(y_i|\boldsymbol{\theta}_k)}{\sum_{k' \in K} \lambda_{k'} f(y_i|\boldsymbol{\theta}_{k'})}. \tag{3.22}$$

$\xi_{i,k}$ is, hence, an observation-specific measure of component-membership, which can also be interpreted as the degree to which observation y_i is consistent with component-model k. Note, however, that the mixture model as defined in (3.1) is based on the assumption that each observation exclusively belongs to one component. The conditional probabilities, thus, reflect an uncertainty about component membership and not some kind of true partial subgroup-membership. An estimator of $\xi_{i,k}$ is given by $\hat{\xi}_{i,k}$.

$\xi_{i,k}$ can be used to classify the observations into K hard clusters: A common allocation rule denoted as Bayes' rule (McLachlan and Peel, 2000, p. 31), maximum a posterior estimator (Gormley and Murphy, 2011, p. 104) or naïve Bayes' classifier (Frühwirth-Schnatter, 2006, p. 27) is to assign each observation to the component to which it most probably belongs to considering $\xi_{i,k}$. Thus, the component-label vector $\boldsymbol{z}_i$ is estimated by $\tilde{\boldsymbol{z}}_i = (\tilde{z}_{i1}, \ldots, \tilde{z}_{iK})^T$, where

$$\tilde{z}_{ik} = \begin{cases} 1 & \text{if } \operatorname{argmax}_k(\xi_{i,k}) \\ 0 & \text{otherwise.} \end{cases} \tag{3.23}$$

This rule is optimal in the sense that it minimizes the expected risk of misclassification under a simple 0/1 loss function that assumes no cost for a correct assignment and equal cost for all possible misallocations (Frühwirth-Schnatter (2006, Chapter 7.1.7), McLachlan and Peel (2000, Chapter 1.15), also see the seminal paper on a decision-theoretic approach to clustering by Binder (1978)). Obviously, $\xi_{i,k}$ is unknown and has to be estimated from the data. Let $\hat{\boldsymbol{\psi}}$ be an estimate of the vector of model parameters and $\hat{\xi}_{i,k} = Pr(z_{ik} = 1|y_i, \hat{\boldsymbol{\psi}})$ be the estimated posterior probability of class-membership, obtained by replacing $\boldsymbol{\psi}$ in (3.22) by $\hat{\boldsymbol{\psi}}$. Then commonly the so-called plug-in version of Bayes' rule (McLachlan and Peel, 2000, p. 31) is applied, which is derived by simply considering $\operatorname{argmax}_k(\hat{\xi}_{i,k})$ instead of $\operatorname{argmax}_k(\xi_{i,k})$ in (3.23).

Chapter 4

Finite Mixture Models for Small Area Estimation

4.1 Introduction

As described in Chapter 2, model-based small area estimators use an explicit statistical model to improve estimation by exploiting the relationship between a set of covariates and the statistic of interest. In some applications, it may, however, be plausible to assume that this relationship differs between different types of units or areas. Given a large enough number of observations, it may then be plausible to estimate different models for different subgroups of units or areas. Then next the questions arises of how to compose sensible subgroups. If no appropriate natural clustering variable is available, FMM might be a suitable framework to "let the data decide" on how to partition the observations into K subgroups and to simultaneously estimate the K subgroup-specific models. In the following sections corresponding mixture-based estimators for SAE are proposed. The predominant aim is to improve the estimation of the statistic of interest by employing a better-fitting model in case of unobserved heterogeneity. Additionally, the fuzzy or hard allocation of observations to subgroups derived as a by-product of the estimation process can provide valuable insights into underlying structures.

The suggested models are further extended to include covariates for the mixture weights. See Section 3.3.3 where the FMM with model for the mixture weights has been introduced in a general framework. As already sketched in the review provided there, this supports the classification of observations when the span of covariates differs between the components. It also provides valuable insights into the clustered structure of the data, thereby intuitively supporting the decision of apply-

ing a mixture-based approach. It further allows to predict subgroup-membership for unsampled areas. Additionally to these advantages, that are also relevant in the general context, in an SAE-application, it can also be expected that the improved assignment to subgroups enhances estimation accuracy for the statistic of interest. Furthermore, the option of assigning new observations to subgroups based on the estimated submodel and the covariates only, can be employed to predict the statistic of interest for unsampled areas in a heterogeneous population.

The novel approach of accounting for the existence of unobserved subgroups of areas via finite mixture models in SAE can draw on literature on mixtures of mixed effects model. Thereby it relies on mixed model and finite mixture model theory as covered in Chapter 2 and 3, respectively. More specifically, in the following sections, the mixture-based small area models are presented as special cases of Finite Mixtures of Mixed Models. Thereby the framework is transferred into the specific language and notation of SAE. Further, theoretical peculiarities and the specific purpose of SAE are accounted for.

4.2 Mixtures of Small Area Models: Framework and Notation

Mixtures of Small Area Models as considered in this thesis are a special case of Finite Mixtures of LMMs. Models of these class have been introduced as an important extension of mixtures of regression models considered in 3.3.2. They offer the possibility to flexibly model correlation between repeated measurements or units from one cluster and at the same time account for unobserved heterogeneity in the population. As such they have been applied in disciplines such as biology, the health sciences and in marketing research (Celeux et al. (2005); Lenk and DeSarbo (2000); Martella et al. (2011); Martinez et al. (2009); McLachlan et al. (2008); Ng and McLachlan (2014); Ng et al. (2006); Wang et al. (2012); Yau et al. (2003), also see Grün (2008)).

Mixtures have been introduced into the mixed model framework in various ways. See Ng and McLachlan (2014) for a discussion of some approaches. In the present context, a mixture of LMM for clustered data as specified in equation 2.5 from Section 2.3 is of interest, where both the fixed coefficients $\boldsymbol{\beta}$ and the covariance matrices of the random effect and the error term, $\boldsymbol{G}$ and $\boldsymbol{R}_i$, are allowed to differ between the components. Under this model the conditional density of $\boldsymbol{y}_i$

given $\boldsymbol{X}_i$ and $\boldsymbol{U}_i$ is given by

$$f(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{U}_i, \boldsymbol{\psi}) = \sum_{k=1}^{K} \lambda_k f_{\mathcal{N}}(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_k, \boldsymbol{V}_{i,k}), \tag{4.1}$$

with $\boldsymbol{V}_{i,k} = \boldsymbol{U}_i\boldsymbol{G}_k\boldsymbol{U}_i^T + \boldsymbol{R}_{i,k}$.

As outlined in the introduction of this chapter, in the context of this thesis Mixtures of LMM are employed as a device to account for the existence of a specified number of different subgroups in the population, each with a distinct relationship between covariates and response variable. At the same time individual observations, i.e. units, are nested in areas with area-specific deviations from the fixed-effect relationship valid in the subgroups (which can also be seen as within-area correlation):

Thus, consider an SAE-setting as introduced in Chapter 2 where n units sampled from a population of size N are nested in m areas, $i = 1, \ldots, m$. Let $\boldsymbol{y}_i$ denote the $n_i \times 1$ of observations $y_{ij}, j = 1, \ldots, n_i$ for area i, which is related to a set of covariates $\boldsymbol{X}_i$ through a linear model. Now, additionally assume that the population is segmented into K non-overlapping, latent subgroups $k = 1, \ldots, K$, also called classes or components. In each subgroup there is a specific relationship between the variable of interest and auxiliary variables. The notational framework is completed by introducing a latent K-dimensional binary component-label vector indicating subgroup membership for area i (see Section 3.3.1). It is $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$, with kth element $z_{ik} = 1$ if area i belongs to class k and $z_{ik} = 0$ otherwise.

Note that two different cases might be of relevance:

- Case 1: *Heterogeneity on area-level*
  The component membership of units in an area is constant. All units in an area share the same fixed and random effects. There are K subgroups of areas.

- Case 2: *Heterogeneity on unit-level*
  The component membership varies between units, regardless of area membership. Units in an area might have different random and fixed effects but share effects with units from other areas. There are K subgroups of units.

Both scenarios can be accounted for by assuming a mixture of K LMM as specified in equation 2.5 from Section 2.3. In the following, the assumption of constant subgroup membership for all units in an area i is drawn, i.e. a case 1-scenario is considered. Note that for this scenario of similarity at the unit level, $\boldsymbol{\beta}_k$ and $\boldsymbol{V}_{i,k}$ are constant for all units within an area.

4.3 Parameter estimation

The log-likelihood function for a mixture of mixed model is straightforwardly obtained from the results presented in Sections 2.3, 3.5 and 4.2 by plugging the density of the component models $f(\boldsymbol{y}_i|\boldsymbol{\theta}_k) = f_{\mathcal{N}}(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_k, \boldsymbol{V}_{i,k})$ into the log-likelihood-function for a general FMM. Thus,

$$l(\boldsymbol{\psi}) = \sum_{i=1}^{m} \log \left(\sum_{k=1}^{K} \lambda_k \frac{\exp\left(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_k)^T \boldsymbol{V}_{i,k}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_k)\right)}{(2\pi)^{n_i/2}(|\boldsymbol{V}_{i,k}|)^{1/2}} \right). \tag{4.2}$$

As common in the mixture model framework, the estimation problem is solved via the EM algorithm (see Section 3.5 and Appendix A.1). Generally, there are two possible approaches, differing with respect to the formulation of the complete data vector. In a first version of the EM algorithm for mixtures of mixed models, both the random effects and the class membership are treated as missing. In this thesis, this approach is taken for the mixture of unit-level models. Alternatively, as in the general EM algorithm for mixture models, it is also possible to only consider the variable indicating class membership as missing information. The finite mixtures of area-level models considered in this work are estimated employing this second strategy. Based on the accounts of the EM algorithm for mixed models described in Section 2.3.5 and for standard finite mixture models discussed in Section 3.5, now respective details for both versions are given. See Celeux et al. (2005), Grün (2008) and Ng and McLachlan (2014), who also provide presentations of the EM algorithm for mixtures of mixed models, partly with slightly different specifications of the underlying model.

4.3.1 Version 1

In the most common version of the EM algorithm for mixtures of mixed models, both the random effects and the class membership are treated as missing. The complete data vector is thus given by $(\boldsymbol{y}^T, \boldsymbol{v}^T, \boldsymbol{z}^T)^T$, where the missing data vector has the form $(\boldsymbol{v}^T, \boldsymbol{z}^T)^T$. Using results from Section 2.3.5 and 3.5 the complete-data

log-likelihood is given by:

$$l_c(\boldsymbol{\psi}) = \sum_{i=1}^{m}\sum_{k=1}^{K} z_{ik}(\log \lambda_k + \log f\left(\boldsymbol{y}_i, \boldsymbol{v}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_k\right)) \tag{4.3}$$

$$= \sum_{i=1}^{m}\sum_{k=1}^{K} z_{ik} \log \lambda_k + \tag{4.4}$$

$$\sum_{i=1}^{m}\sum_{k=1}^{K} z_{ik}\left(-\frac{1}{2}\left(\boldsymbol{d}_i^T \boldsymbol{\Sigma}_i^{-1}\boldsymbol{d}_i + \log|\boldsymbol{\Sigma}_i| + (n_i+s)\log(2\pi)\right)\right),$$

where

$$\boldsymbol{d}_i = \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{X_i}\boldsymbol{\beta} \\ \boldsymbol{v}_i - \boldsymbol{0} \end{pmatrix}, \tag{4.5}$$

and

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{U}_i\boldsymbol{G}\boldsymbol{U}_i^T + \sigma^2\boldsymbol{I}_{n_i} & \boldsymbol{U}_i\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{U}_i^T & \boldsymbol{G} \end{pmatrix} \tag{4.6}$$

being the variance covariance matrix of the joint distribution of $(\boldsymbol{y}_i^T, \boldsymbol{v}_i^T)^T$.

- *E-step:*
  In the *E-step* the conditional expectation of the complete-data log-likelihood $Q(\boldsymbol{\psi}; \widehat{\boldsymbol{\psi}}^{(t-1)}) = E_{\widehat{\boldsymbol{\psi}}^{t-1}}[l_c(\boldsymbol{\psi}|\boldsymbol{y})]$ is obtained. It can be seen from (4.3) that this requires deriving the following conditional moments of the missing data: The conditional expectation for class membership, i.e. $\xi_{i,k}$ (compare Section 3.5), and the conditional expectation of the sufficient statistics $\boldsymbol{v}_i$ and $\boldsymbol{v}_i\boldsymbol{v}_i^T$ (compare Section 2.3.5). It is

$$\hat{\xi}_{i,k}^{(t-1)} = Pr_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(z_{ik} = 1|\boldsymbol{y}_i) \tag{4.7}$$

$$= \frac{\hat{\lambda}_k^{(t-1)} f_{\mathcal{N}}(\boldsymbol{y}_i|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_k^{(t-1)})}{\sum_{j\in K}\hat{\lambda}_j^{(t-1)} f_{\mathcal{N}}(\boldsymbol{y}_i|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_j^{(t-1)})}$$

$$\hat{\boldsymbol{s}}_{1i,k}^{(t-1)} = E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(\boldsymbol{v}_i|\boldsymbol{y}_i) \tag{4.8}$$

$$= (\boldsymbol{U}_i^T\boldsymbol{U}_i + \hat{\sigma}_{e,k}^{2(t-1)}\widehat{\boldsymbol{G}}^{(t-1)^{-1}})^{-1}\boldsymbol{U}_i^T(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_k^{(t-1)})$$

$$\widehat{\boldsymbol{S}}_{2i,k}^{(t-1)} = E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(\boldsymbol{v}_i\boldsymbol{v}_i^T|\boldsymbol{y}_i) \tag{4.9}$$

$$= \mathrm{Cov}_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(\boldsymbol{v}_i|\boldsymbol{y}_i) + \hat{\boldsymbol{s}}_{1i,k}^{(t-1)}\hat{\boldsymbol{s}}_{1i,k}^{(t-1)^T}$$

$$= ((\hat{\sigma}_{e,k}^{2(t-1)})^{-1}\boldsymbol{U}_i^T\boldsymbol{U}_i + \widehat{\boldsymbol{G}}^{(t-1)^{-1}})^{-1} + \widehat{\boldsymbol{s}}_{1i,k}^{(t-1)}\widehat{\boldsymbol{s}}_{1i,k}^{(t-1)^T},$$

- *M-step*:
  It is obvious from the complete-data log-likelihood (4.3), that optimization for the mixture weights (or the submodel for the mixture weights in case of a mixture with concomitant variables) and the parameters in the main model can be done separately, i.e. estimation of λ_k (or $\boldsymbol{\alpha}$) is independent of the form of the component densities. Therefore, for these parameters results from the general representation of the EM algorithm for mixture models in 3.5 apply.

  As there are no parameters that are fixed over the components, estimation of parameters in the main model can be performed for each component separately. Maximizing Q with respect to the component-specific parameters yields:

$$\widehat{\boldsymbol{\beta}}_k^{(t)} = \frac{1}{\sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)}} (\sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)} \boldsymbol{X}_i^T \boldsymbol{X}_i)^{-1} \sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)} \boldsymbol{X}_i^T (\boldsymbol{y}_i - \boldsymbol{U}_i \hat{\boldsymbol{s}}_{1i,k}^{(t-1)}), \quad (4.10)$$

$$\widehat{\boldsymbol{G}}_k^{(t)} = \frac{1}{\sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)}} \sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)} \widehat{\boldsymbol{S}}_{2i}^{(t-1)}, \quad (4.11)$$

$$\hat{\sigma}_{e,k}^{2(t)} = \frac{1}{\sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)} n_i} \sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)} [(\boldsymbol{\epsilon}_{i,k}^{(t-1)^T} \boldsymbol{\epsilon}_{i,k}^{(t-1)}), \quad (4.12)$$

$$+ \operatorname{tr} \left(\boldsymbol{U}_i^T \boldsymbol{U}_i \left((\hat{\sigma}_{e,k}^{2(t-1)})^{-1} \boldsymbol{U}_i^T \boldsymbol{U}_i + \widehat{\boldsymbol{G}}^{(t-1)^{-1}} \right)^{-1} \right)], \quad (4.13)$$

  where

$$\hat{\boldsymbol{\epsilon}}_{i,k}^{(t-1)} = \boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_k^{(t-1)} - \boldsymbol{U}_i \hat{\boldsymbol{s}}_{1i,k}^{(t-1)}. \quad (4.14)$$

4.3.2 Version 2

Alternatively, a straightforward application of the EM algorithm for FMM as presented in Section 3.5 can be applied (GRÜN, 2008). In this framework only the subgroup membership is treated as missing. As described in the presentation of the *M-step* in Section 3.5, this requires maximizing the weighted log-likelihood for each of the mixture components. As GRÜN (2008) points out, software for weighted maximum likelihood estimation for linear mixed models that also allows for varying covariance matrices for the error term is, however, not readily available. For certain simple model specifications, she, therefore, proposes a transformation to the data which allows to maximize the unweighted log-likelihood instead of the weighted one: If $\boldsymbol{U}_i = \boldsymbol{U}$ for all $i = 1, \ldots, m$, the solution to unweighted ML estimation for the transformed data $\check{\boldsymbol{y}} = \sqrt{\boldsymbol{\xi}_i} \boldsymbol{y}_i$ and $\check{\boldsymbol{x}} = \sqrt{\boldsymbol{\xi}_i} \boldsymbol{x}_i$,

where $\boldsymbol{\xi}_i = (\xi_{i,1}, \ldots, \xi_{i,K})^T$, is equal to the results obtained from weighted ML on the original data. See Grün (2008) for proofs and further details. In the context of this thesis, the function `maxLik` from the `maxLik` package (Henningsen and Toomet, 2011) was employed in order to solve the maximization problem numerically.

4.4 Prediction from Mixtures of Small Area Models

As described in Section 2.3.3, in the mixed model context commonly prediction of a general linear combination $\eta_i = \boldsymbol{l}_i^T\boldsymbol{\beta} + \boldsymbol{m}_i^T\boldsymbol{v}_i$ involving both fixed and random effects is of interest. Relying on the concepts and notation introduced there, now a corresponding predictor under a mixture of LMM is developed.

Under the drawn assumption of a population segmented into K disjoint subgroups with component-specific model parameters, for an area i belonging to component k, η_i is given by $\eta_i|(z_{ik} = 1) = \eta_{ik} = \boldsymbol{l}_i^T\boldsymbol{\beta}_k + \boldsymbol{m}_i^T\boldsymbol{v}_{i,k}$. Introducing $\eta_{ik}^* = \boldsymbol{l}_i^T\boldsymbol{\beta}_k + \boldsymbol{m}_i^T\boldsymbol{v}_{i,k}, k = 1, \ldots, K$, the true value η_i for areas from cluster 1 to K can, thus, be written more generally as

$$\begin{aligned} \eta_i &= \sum_{k=1}^{K} z_{ik}\eta_{ik}^* \\ &= \sum_{k=1}^{K} z_{ik}(\boldsymbol{l}_i^T\boldsymbol{\beta}_k + \boldsymbol{m}_i^T\boldsymbol{v}_{i,k}) \end{aligned} \tag{4.15}$$

Note that for an area belonging to k, η_{ik}^* is the true value of the mixed effect η_{ik}, while for all other components it is a theoretical construct without meaningful interpretation, introduced for the sake of a compact representation of η_i valid for areas from all components. This representation will be drawn upon in developing a predictor for η_i.

As in the standard framework of SAE, interest is in a prediction that takes the realized value of the random effect into account (see 2.3.3). Further, in the introduced setting of a segmented population the individual subgroup membership should be factored in, i.e. the realized value of the multinomial variable indicating cluster membership is of interest, too. Thus, focus is on a conditional prediction instead of a marginal prediction that averages over the unobserved random variables $\boldsymbol{v}$ and $\boldsymbol{z}$. To date, there seems to be no systematic treatment of conditional

or individual prediction from mixture models in the literature. See Cole and Bauer (2016) for the only recent exemption.

The best predictor for a random variable is its conditional expectation given the observed data. Assuming all model parameters are known, the conditional expectation of η_i given $\boldsymbol{y}_i$ and $\boldsymbol{l}_i$ is

$$\begin{aligned} E(\eta_i|\boldsymbol{y}_i, \boldsymbol{l}_i) &= E(\sum_{k=1}^{K} z_{ik}\eta_{ik}^*|\boldsymbol{y}_i, \boldsymbol{l}_i) \\ &= \sum_{k=1}^{K} E(z_{ik}|\boldsymbol{y}_i, \boldsymbol{l}_i) \cdot E(\eta_{ik}^*|\boldsymbol{y}_i, \boldsymbol{l}_i) + \text{Cov}(z_{ik}\eta_{ik}^*|\boldsymbol{y}_i, \boldsymbol{l}_i) \end{aligned} \tag{4.16}$$

$\eta_{ik}^* = \boldsymbol{l}_i^T\boldsymbol{\beta}_k + \boldsymbol{m}_i^T\boldsymbol{v}_{i,k}$, so that $\eta_{ik}^*|\boldsymbol{l}_i$ varies only by $\boldsymbol{m}_i^T\boldsymbol{v}_{i,k}$. Thus, assuming independence between the random effects and $\boldsymbol{z}_i$ implies that $\text{Cov}(z_{ik}\eta_{ik}^*|\boldsymbol{y}_i, \boldsymbol{l}_i) = 0$.

It is (compare Section 2.3.3)

$$E(\eta_{ik}^*|\boldsymbol{y}_i, \boldsymbol{l}_i) = \tilde{\eta}_{ik}^* = \boldsymbol{l}_i^T\boldsymbol{\beta}_k + \boldsymbol{m}_i^T\tilde{\boldsymbol{v}}_{i,k}, \tag{4.17}$$

with

$$\tilde{\boldsymbol{v}}_{i,k} = E(\boldsymbol{v}_{i,k}|\boldsymbol{y}_i, \boldsymbol{l_i}) = \boldsymbol{G}_k\boldsymbol{U}_i^T\boldsymbol{V}_{i,k}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_k) \tag{4.18}$$

and (compare Section 3.7)

$$E(z_{ik}|\boldsymbol{y}_i, \boldsymbol{l}_i) = Pr(z_{ik} = 1|\boldsymbol{y}_i, \boldsymbol{l}_i) = \xi_{i,k}, \tag{4.19}$$

where

$$\xi_{i,k} = \frac{\lambda_k f_{\mathcal{N}}(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_k, \boldsymbol{V}_{i,k})}{\sum_{k'=1}^{K} \lambda_{k'} f_{\mathcal{N}}(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_{k'}, \boldsymbol{V}_{i,k'})}, \tag{4.20}$$

which is the posterior or conditional probability that area i belongs to class k already introduced in Section 3.7 and in the context of the EM algorithm for mixture models in Section 3.5. Thus,

$$E(\eta_i|\boldsymbol{y}_i, \boldsymbol{l}_i) = \sum_{k=1}^{K} \xi_{i,k}\tilde{\eta}_{ik}^*. \tag{4.21}$$

Plugging in the ML-estimates of the model parameters yields the following predictor for η_i that accounts for the existence of unobserved subgroups of areas:

$$\hat{\eta}_i^{\text{mix}} = \sum_{k=1}^{K} \hat{\xi}_{i,k}\hat{\eta}_{ik}^* \tag{4.22}$$

$$= \sum_{k=1}^{K} \hat{\xi}_{i,k}(\boldsymbol{l}_i^T\hat{\boldsymbol{\beta}}_k + \boldsymbol{m}_i^T\hat{\boldsymbol{v}}_{i,k}), \tag{4.23}$$

where $\hat{\eta}_i^*$ and $\hat{\xi}_{i,k}$ are derived from $\tilde{\eta}_i^*$ and $\xi_{i,k}$, respectively, by replacing the true parameter values $\boldsymbol{\psi}$ by $\widehat{\boldsymbol{\psi}}$. Thus, η_i is estimated as a weighted mean of predicts from the K models, where the area-specific weights are given by the conditional probabilities that area i belongs to model k.

If a submodel for the mixture model is applied, respective results are obtained accordingly by additionally conditioning on $\boldsymbol{w}_i$, i.e. the covariates in the submodel:

$$E(\eta_i|\boldsymbol{y}_i, \boldsymbol{l}_i, \boldsymbol{w}_i) = \sum_{k=1}^{K} \xi_{i,k}^c \tilde{\eta}_{ik}^*, \tag{4.24}$$

where

$$\xi_{i,k}^c = \frac{\lambda_k(\boldsymbol{w}_i, \boldsymbol{\alpha}_k) f_{\mathcal{N}}(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_k, \boldsymbol{V}_{i,k})}{\sum_{k'=1}^{K} \lambda_{k'}(\boldsymbol{w}_i, \boldsymbol{\alpha}_{k'}) f_{\mathcal{N}}(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_{k'}, \boldsymbol{V}_{i,k'})}. \tag{4.25}$$

Plugging in estimated model parameters yields

$$\eta_i^{\text{mix,c}} = \sum_{k=1}^{K} \hat{\xi}_{i,k}^c \hat{\eta}_{ik}^* \tag{4.26}$$

as before, only with the modified expression for $\xi_{i,k}$.

Applying a model-based approach to small area estimation, opens up the option of calculating estimates for unsampled areas, i.e. areas for which no direct estimate can be obtained for the statistic of interest. In what follows a predictor for out-of-sample prediction is derived. This is especially promising if a mixture model with submodel for the mixture weights is considered because it allows to predict an area-specific mixture weight from the submodel based on the covariates only.

Starting again from (4.15) and taking the expectation of η_i given $\boldsymbol{l}_i$ and the covariates in the submodel, $\boldsymbol{w}_i$, yields

$$\begin{aligned} E(\eta_i|\boldsymbol{l}_i, \boldsymbol{w}_i) &= E(\sum_{k=1}^{K} z_{ik}\eta_{ik}^*|\boldsymbol{l}_i, \boldsymbol{w}_i) \\ &= \sum_{k=1}^{K} \lambda_{i,k} \boldsymbol{l}_i^T \boldsymbol{\beta}_k. \end{aligned} \tag{4.27}$$

Note that $\lambda_{i,k} = \lambda_k$ if no submodel for the mixture weights is assumed.

As before, a predictor is derived by replacing the true model parameters $\boldsymbol{\psi}$ by their ML estimate $\widehat{\boldsymbol{\psi}}$:

$$\hat{\eta}_i^{\text{mix.oos}} = \sum_{k=1}^{K} \hat{\lambda}_{i,k} \boldsymbol{l}_i^T \hat{\boldsymbol{\beta}}_k. \tag{4.28}$$

Based on the estimated model and available covariates $\boldsymbol{w}$ and $\boldsymbol{l}$, an out-of-sample prediction for the statistic of interest is, thus, derived as a weighted mean of synthetic estimates from the component models where individual weights are predicted from the submodel.

4.5 A Mixture of Area-level Models

Similar to the approach taken in Chapter 2, where the Fay-Herriot model was introduced as a special case of a general LMM, in this section a finite mixture of area-level models is presented as a special case of the Finite Mixture of LMMs.

Starting from the framework introduced in Section 2.4, now consider a scenario where the m areas are divided into K disjoint classes, each with a specific relationship between response variable and covariates. Extending the setting from the standard model, it is therefore assumed that the observed direct estimate for a given area i from subgroup k is appropriately modelled by a Fay-Herriot-type LMM with component-specific fixed coefficients $\boldsymbol{\beta}_k$ and model variance $\sigma^2_{v,k}$, i.e. $\hat{\mu}_i^{\text{Dir}}|(z_{ik} = 1) = \mu_{ik} + e_i$, where $\mu_{ik} = \mu_i|(z_{ik} = 1) = \boldsymbol{x}_i^T\boldsymbol{\beta}_k + v_{i,k}$ with $v_{i,k} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2_{v,k})$ denotes the true area mean for areas in k. As before, e_i is the sampling error following a normal distribution with known design variance $\sigma^2_{e,i}$. Thus, the conditional distribution of $\hat{\mu}_i^{\text{Dir}}$ given $z_{ik} = 1$ is a univariate normal distribution characterized by the density $f_{\mathcal{N}}(\hat{\mu}_i^{\text{Dir}}; \boldsymbol{x}_i^T\boldsymbol{\beta}_k, \sigma^2_{e,i} + \sigma^2_{v,k})$.

The marginal density of $\hat{\mu}_i^{\text{Dir}}$ then is a finite mixture of K Fay-Herriot-type LMM, i.e.

$$f(\hat{\mu}_i^{\text{Dir}}) = \sum_{k=1}^{K} \lambda_k f_{\mathcal{N}}(\hat{\mu}_i^{\text{Dir}}; \boldsymbol{x}_i^T\boldsymbol{\beta}_k, \sigma^2_{e,i} + \sigma^2_{v,k}), \quad i = 1, \ldots, m. \tag{4.29}$$

This model is a special case of the finite mixture of LMM obtained from (4.1) by making the replacements given in (2.49) and (2.50), only that now $\boldsymbol{v}_{i,k} = v_{i,k}$, $\boldsymbol{G}_k = \sigma^2_{v,k}$ and $\boldsymbol{V}_{i,k} = \sigma^2_{e,i} + \sigma^2_{v,k}$.

Estimates for the component-specific parameters $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T, \sigma^2_{v,1}, \ldots, \sigma^2_{v,K})^T$ and the mixing proportions $\lambda_1, \ldots, \lambda_K$ are obtained via the EM algorithm for mixtures of LMM. In Section 4.3 two versions of the algorithm were described. In what follows this general account is complemented by providing details for the specific case of a mixture of area-level models.

For the first version, commonly applied for the estimation of mixtures of LMM, both the component membership and the random effects are treated as missing.

Expressions for the *E-step* and the *M-step* for the specific case of the mixture of area-level models are as follows:

- *E-step*
  An updated estimate for the conditional expectation of subgroup membership, $\xi_{i,k}$, is obtained as

$$\hat{\xi}_{i,k}^{(t-1)} = \frac{\hat{\lambda}_k^{(t-1)} f_{\mathcal{N}}(\hat{\mu}_i^{\text{Dir}} | \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_k^{(t-1)})}{\sum_{j \in K} \hat{\lambda}_j^{(t-1)} f_{\mathcal{N}}(\hat{\mu}_i^{\text{Dir}} | \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_j^{(t-1)})} \tag{4.30}$$

  The expressions for the conditional moments of the random effect in the *E-step* simplify to:

$$\begin{aligned} \hat{s}_{1i,k}^{(t-1)} &= E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(v_{i,k} | \hat{\mu}_i^{\text{Dir}}) && (4.31) \\ &= \frac{\hat{\sigma}_{v,k}^{2^{(t-1)}}}{\sigma_{e,i}^2 + \hat{\sigma}_{v,k}^{2^{(t-1)}}} (\hat{\mu}_i^{\text{Dir}} - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_k^{(t-1)}) \\ \hat{S}_{2i,k}^{(t-1)} &= E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(v_{i,k}^2 | \hat{\mu}_i^{\text{Dir}}) && (4.32) \\ &= \frac{\sigma_{e,i}^2 \hat{\sigma}_{v,k}^{2^{(t-1)}}}{\sigma_{e,i}^2 + \hat{\sigma}_{v,k}^{2^{(t-1)}}} + \hat{s}_{1i,k}^{(t-1)} \hat{s}_{1i,k}^{(t-1)} \end{aligned}$$

- *M-step*
  The expressions for deriving an updated fit for the model parameters in the *M-step* are obtained as:

$$\widehat{\boldsymbol{\beta}}_k^{(t)} = \frac{1}{\sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)}} (\sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)} \boldsymbol{x}_i \boldsymbol{x}_i^T)^{-1} \sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)} \boldsymbol{x}_i (\hat{\mu}_i^{\text{Dir}} - \hat{s}_{1i,k}^{(t-1)}). \tag{4.33}$$

$$\hat{\sigma}_{v,k}^{2(t)} = \frac{1}{\sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)}} \sum_{i=1}^m \hat{\xi}_{i,k}^{(t-1)} \widehat{S}_{2i}^{(t-1)} \tag{4.34}$$

With the second version an alternative algorithm was presented, where only the component membership is treated as missing. This requires the maximization of the weighted log-likelihood of a mixed model in the *M-step*. For the mixture of area-level models, this second approach was taken and the maximization problem was solved numerically using `maxLik` from the `maxLik` package (Henningsen and Toomet, 2011).

Finally, the mixture-based estimator for the true area mean is derived from the general predictor developed in Section 4.4. Corresponding to the approach taken

in the preceding section, introduce $\mu_i = \sum_{k=1}^{K} z_{ik}\mu_{ik}^*$, where $\mu_{ik}^* = (\boldsymbol{x}_i^T\boldsymbol{\beta}_k + v_{i,k})$, $i = 1, \ldots, K$. Note that μ_i under the mixture model, thus, is a special case of η_i as considered in 4.4 (also compare Section 2.4). Making the relevant replacements in (4.22) yields

$$\hat{\mu}_i^{\text{FHmix}} = \sum_{k=1}^{K} \hat{\xi}_{i,k}\hat{\mu}_{ik}^{*\text{FH}} \tag{4.35}$$

$$= \sum_{k=1}^{K} \hat{\xi}_{i,k}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_k + \hat{\gamma}_{i,k}(\hat{\mu}_i^{\text{Dir}} - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_k), \tag{4.36}$$

with

$$\hat{\gamma}_{i,k} = \frac{\hat{\sigma}_{v,k}^2}{\sigma_{e,i}^2 + \hat{\sigma}_{v,k}^2}. \tag{4.37}$$

$\hat{\mu}_{ik}^{*\text{FH}}$ denotes the predict for μ_i derived from model k.

The mixture-based estimator can be rearranged as follows:

$$\begin{aligned}
\hat{\mu}_i^{\text{FHmix}} &= \sum_{k=1}^{K} \hat{\xi}_{i,k} \cdot \hat{\mu}_{ik}^{*\text{FH}} \\
&= \sum_{k=1}^{K} \hat{\xi}_{i,k}\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_k + \sum_{k=1}^{K} \hat{\xi}_{i,k}\hat{v}_{i,k} \\
&= \underbrace{\sum_{k=1}^{K} \hat{\xi}_{i,k}\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_k}_{\text{Synthetic estimator}} + \underbrace{\sum_{k=1}^{K} \hat{\xi}_{i,k}\hat{\gamma}_{i,k}(\hat{\mu}_i^{Dir} - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_k)}_{\text{Correction of synthetic estimator}}.
\end{aligned} \tag{4.38}$$

The FHmix-estimator, thus, can be formulated as a synthetic estimator obtained as the weighted mean of predicts from the fixed parts of the K models and an area-specific correction factor, that adjusts the synthetic estimator. This correction factor is a convex combination of component-specific corrections.

Second, the FHmix-estimator can simply be interpreted as a convex combination of K composite estimators.

$$\begin{aligned}
\hat{\mu}_i^{\text{FHmix}} &= \sum_{k=1}^{K} \hat{\xi}_{i,k} \cdot \hat{\mu}_{ik}^{*\text{FH}} \\
&= \sum_{k=1}^{K} \hat{\xi}_{i,k} \left(\hat{\gamma}_{i,k}\hat{\mu}_i^{\text{Dir}} + (1 - \hat{\gamma}_{i,k})\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_k \right)
\end{aligned} \tag{4.39}$$

Further pursuing the notion of a composite estimator that yields an area-specific compromise between the influence of the direct estimator and the synthetic estimator, the FHmix-estimator can finally be interpreted the following way:

$$\hat{\mu}_i^{\text{FHmix}} = \sum_{k=1}^{K} \hat{\xi}_{i,k} \left(\hat{\gamma}_{i,k} \hat{\mu}_i^{\text{Dir}} \right) + \sum_{k=1}^{K} \hat{\xi}_{i,k} \left((1 - \hat{\gamma}_{i,k}) \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_k \right). \tag{4.40}$$

It can, thus, be written as the sum of (1) a convex combination of contributions from the direct estimator in the predicts from model 1 to K and (2) the respective weighted sum of contributions of the synthetic estimator.

4.6 A Mixture of Unit-level Models

Corresponding to the approach taken in the preceding section, now a mixture of unit-level models is introduced as a special case of the mixture of LMMs.

Consider the small area scenario as defined in Section 2.5. Now, as described in Section 4.2 the framework is extended by assuming the existence of K latent classes of areas, each with distinct model parameters $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T, \sigma_{v,k}^2, \sigma_{e,k}^2)^T$. Thus, for a given area i in subgroup k the relevant model that links the unit-level observations of the statistic of interest to a set of auxiliary information is given by

$$\begin{aligned} y_{ij}|(z_{ik} = 1) &= \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k + v_{i,k} + e_{ij,k}, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, n_i \\ v_{i,k} &\overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_{v,k}^2) \\ e_{ij,k} &\overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_{e,k}^2). \end{aligned} \tag{4.41}$$

Under this model, the conditional density of the statistic of interest given $z_{ik} = 1$ is $f_{\mathcal{N}}(\boldsymbol{y}_i; \boldsymbol{X}_i \boldsymbol{\beta}_k, \sigma_{e,k}^2 \boldsymbol{I}_{n_i} + \sigma_{v,k}^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T)$ and the marginal density is given by the finite mixture

$$f(\boldsymbol{y}_i | \boldsymbol{X}_i, \boldsymbol{U}_i, \boldsymbol{\psi}) = \sum_{k=1}^{K} \lambda_k f_{\mathcal{N}}(\boldsymbol{y}_i; \boldsymbol{X}_i \boldsymbol{\beta}_k, \sigma_{e,k}^2 \boldsymbol{I}_{n_i} + \sigma_{v,k}^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T), \tag{4.42}$$

which is a special case of (4.1) with covariance matrix $\boldsymbol{V}_{i,k} = \sigma_{e,k}^2 \boldsymbol{I}_{n_i} + \sigma_{v,k}^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T$.

As described in Section 4.3, ML estimates for this model class are commonly derived using the EM algorithm and treating both the random effects and the subgroup membership as missing (Version 1). This approach was adopted for the

estimation of mixtures of unit-level models in this thesis. Corresponding equations are derived straightforwardly from the general equations by setting $\boldsymbol{U}_i = \mathbf{1}_{n_i}$, $\boldsymbol{G}_k = \sigma^2_{v,k}$ and $\boldsymbol{R}_i = \boldsymbol{I}_{n_i}$ (compare Section 2.5):

- *E-step*
  Expressions for the conditional moments in the *E-step* simplify to

$$\begin{aligned} \hat{s}^{(t-1)}_{1i,k} &= E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(v_i|\boldsymbol{y}_i) \\ &= \hat{\gamma}^{(t-1)}_{i,k}(\bar{y}_i - \bar{\boldsymbol{x}}^T_{iS}\hat{\boldsymbol{\beta}}^{(t-1)}_k), \end{aligned} \tag{4.43}$$

  where

$$\hat{\gamma}^{(t-1)}_{i,k} = \frac{\hat{\sigma}^{2^{(t-1)}}_{v,k}}{\hat{\sigma}^{2^{(t-1)}}_{v,k} + \frac{\hat{\sigma}^{2^{(t-1)}}_{e,k}}{n_i}}, \tag{4.44}$$

  and

$$\begin{aligned} \widehat{S}^{(t-1)}_{2i,k} &= E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}(v_i v_i^T|\boldsymbol{y}_i) \\ &= \left(\frac{1}{\hat{\sigma}^{2^{(t-1)}}_{e,k}} n_i + \frac{1}{\hat{\sigma}^{2^{(t-1)}}_{v,k}}\right)^{-1} + \hat{s}^{(t-1)}_{1i,k}\hat{s}^{(t-1)}_{1i,k}, \\ &= \frac{1}{n_i}\frac{\hat{\sigma}^{2^{(t-1)}}_{e,k}\hat{\sigma}^{2^{(t-1)}}_{v,k}}{\hat{\sigma}^{2^{(t-1)}}_{v,k} + \frac{\hat{\sigma}^{2^{(t-1)}}_{e,k}}{n_i}} + \hat{s}^{(t-1)}_{1i,k}\hat{s}^{(t-1)}_{1i,k}. \end{aligned} \tag{4.45}$$

- *M-step*
  Updated estimates in the *M-step* are obtained as

$$\widehat{\boldsymbol{\beta}}^{(t)}_k = \frac{1}{\sum_{i=1}^m \hat{\xi}^{(t-1)}_{i,k}}\Big(\sum_{i=1}^m \hat{\xi}^{(t-1)}_{i,k}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}\boldsymbol{x}^T_{ij}\Big)^{-1}\sum_{i=1}^m \hat{\xi}^{(t-1)}_{i,k}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}(y_{ij} - \hat{s}^{(t-1)}_{1i,k}), \tag{4.46}$$

$$\hat{\sigma}^{2^{(t)}}_{v,k} = \frac{1}{\sum_{i=1}^m \hat{\xi}^{(t-1)}_{i,k}}\sum_{i=1}^m \hat{\xi}^{(t-1)}_{i,k}\widehat{S}^{(t-1)}_{2i}, \tag{4.47}$$

$$\hat{\sigma}^{2^{(t)}}_{e,k} = \frac{1}{\sum_{i=1}^m \hat{\xi}^{(t-1)}_{i,k} n_i}\sum_{i=1}^m \hat{\xi}^{(t-1)}_{i,k}\left[\sum_{i=1}^{n_i}\epsilon^{(t-1)}_{ij,k}\epsilon^{(t-1)}_{ij,k} + \frac{\hat{\sigma}^{2^{(t-1)}}_{e,k}\hat{\sigma}^{2^{(t-1)}}_{v,k}}{\hat{\sigma}^{2^{(t-1)}}_{v,k} + \frac{\hat{\sigma}^{2^{(t-1)}}_{e,k}}{n_i}}\right], \tag{4.48}$$

  where

$$\hat{\epsilon}^{(t-1)}_{ij,k} = y_{ij} - \boldsymbol{x}^T_i\hat{\boldsymbol{\beta}}^{(t-1)}_k - \hat{s}^{(t-1)}_{1i,k}. \tag{4.49}$$

Once estimates for the model parameters are obtained, the mixture-based predictor for the statistic of interest is derived. As discussed in Section 2.5, commonly the expected value of the area mean is considered as the target statistic. Under the assumed mixture model, for an area belonging to class k it is given by $\mu_i|(z_{ik} = 1) = \bar{\boldsymbol{x}}_{iP}^T\boldsymbol{\beta}_k + v_{i,k}$. As before (compare Section 4.4) define $\mu_i = \sum_{k=1}^K z_{ik}\mu_{ik}^*$, where $\mu_{ik}^* = \bar{\boldsymbol{x}}_{iP}^T\boldsymbol{\beta}_k + v_{i,k}, i = 1, \ldots, K$, which is a special case of η_i under the mixture model as introduced in (4.15). A predictor is derived from the general predictor developed in Section 4.4 by making the relevant replacements in (4.22). It is

$$\hat{\mu}_i^{\text{BHFmix}} = \sum_{k=1}^K \hat{\xi}_{i,k}\hat{\mu}_{ik}^{*\text{BHF}} \tag{4.50}$$

$$= \sum_{k=1}^K \hat{\xi}_{i,k}(\bar{\boldsymbol{x}}_{iP}^T\hat{\boldsymbol{\beta}}_k + \hat{\gamma}_{i,k}(\bar{y}_i - \bar{\boldsymbol{x}}_{iS}^T\hat{\boldsymbol{\beta}}_k), \tag{4.51}$$

with

$$\hat{\gamma}_{i,k} = \frac{\hat{\sigma}_{v,k}^2}{\hat{\sigma}_{v,k}^2 + \frac{\hat{\sigma}_{e,i}^2}{n_i}}. \tag{4.52}$$

$\hat{\mu}_{ik}^{*\text{BHF}}$ denotes the predict for μ_i derived from model k.

4.7 MSE Estimation

A potential measure of uncertainty can be obtained as $\text{MSE}(\hat{\eta}_i^{\text{mix}}|\hat{\eta}_i^{\text{Dir}}) := E[(\hat{\eta}_i^{\text{mix}} - \eta_i)^2|\boldsymbol{y}_i]$, i.e. as the conditional expectation of the MSE given the observed data (see Jiang, 2017, p. 148 and references therein for a similar conditional uncertainty measure for the EBLUP). In what follows, an approximation of $\text{MSE}(\hat{\eta}_i^{\text{mix}}|\hat{\eta}_i^{\text{Dir}})$ for the special case of a mixture of area-level models is presented (Articus and Burgard, forthcoming). To obtain it, the conditional MSE is expressed in a way that allows to approximate its terms by known expressions. More precisely,

$\text{MSE}(\hat{\mu}_i^{\text{mix}}|\hat{\mu}_i^{\text{Dir}}) := E[(\hat{\mu}_i^{\text{mix}} - \mu_i)^2|\hat{\mu}_i^{\text{Dir}}]$ can be transformed as follows:

$$\begin{aligned}
\text{MSE}(\hat{\mu}_i^{\text{mix}}|\hat{\mu}_i^{\text{Dir}}) :=& E[(\hat{\mu}_i^{\text{mix}} - \mu_i)^2|\hat{\mu}_i^{\text{Dir}}] \\
=& E\left[E[(\hat{\mu}_i^{\text{mix}} - \mu_i)^2|\hat{\mu}_i^{\text{Dir}}, \boldsymbol{z}_i]|\hat{\mu}_i^{\text{Dir}}\right] \\
=& \sum_{k=1}^{K} Pr(z_{ik} = 1|\hat{\mu}_i^{\text{Dir}}) \times E\left[(\hat{\mu}_i^{\text{mix}} - \mu_i)^2|\hat{\mu}_i^{\text{Dir}}, z_{ik} = 1\right] \\
=& \sum_{k=1}^{K} \xi_{ik} \times E\left[(\hat{\mu}_i^{\text{mix}} - \mu_{ik}^* + \mu_{ik}^* - \mu_i)^2|\hat{\mu}_i^{\text{Dir}}, z_{ik} = 1\right] \\
=& \sum_{k=1}^{K} \xi_{ik} \times E\big[(\hat{\mu}_i^{\text{mix}} - \mu_{ik}^*)^2 \\
& + 2(\hat{\mu}_i^{\text{mix}} - \mu_{ik}^*)(\mu_{ik}^* - \mu_i) + (\mu_{ik}^* - \mu_i)^2|\hat{\mu}_i^{\text{Dir}}, z_{ik} = 1\big] \\
=& \sum_{k=1}^{K} \xi_{ik} \times \Big(E\left[(\hat{\mu}_i^{\text{mix}} - \mu_{ik}^*)^2|\hat{\mu}_i^{\text{Dir}}, z_{ik} = 1\right] \\
& + 2E\left[(\hat{\mu}_i^{\text{mix}} - \mu_{ik}^*)(\mu_{ik}^* - \mu_i)|\hat{\mu}_i^{\text{Dir}}, z_{ik} = 1\right] \\
& + E\left[(\mu_{ik}^* - \mu_i)^2|\hat{\mu}_i^{\text{Dir}}, z_{ik} = 1\right]\Big)
\end{aligned} \tag{4.53}$$

The following approximations are suggested:

$$E\left[(\hat{\mu}_i^{\text{mix}} - \mu_{ik}^*)^2|\hat{\mu}_i^{\text{Dir}}, z_{ik} = 1\right] \approx \widehat{\text{MSE}}_k \tag{4.54}$$

$$2E\left[(\hat{\mu}_i^{\text{mix}} - \mu_{ik}^*)(\mu_{ik}^* - \mu_i)|\hat{\mu}_i^{\text{Dir}}, z_{ik} = 1\right] \approx 0 \tag{4.55}$$

$$E\left[(\mu_{ik}^* - \mu_i)^2|\hat{\mu}_i^{\text{Dir}}, z_{ik} = 1\right] \approx (\hat{\mu}_{ik}^{*\text{FH}} - \hat{\mu}_i^{\text{mix}})^2 \tag{4.56}$$

Note that (4.54) can be understood as some kind of within error, i.e. the prediction error within a component. It is approximated by $\widehat{\text{MSE}}_k$, which, in accordance to the MSE estimator of the EBLUP for variance components estimated by ML (see Section 2.3.4 and 2.4), is obtained as

$$\begin{aligned}
\widehat{\text{MSE}}_k \approx \quad & g_1(\hat{\sigma}_{v,k}^2) + g_2(\hat{\sigma}_{v,k}^2) + 2g_3(\hat{\sigma}_{v,k}^2) \\
& - \left(\frac{\sigma_{e,i}^2}{\sigma_{v,k}^2 + \sigma_{e,i}^2}\right)^2 \left(\sum_{j=1}^{m} \frac{1}{(\sigma_{v,k}^2 + \sigma_{e,j}^2)^2}\right)^{-1} \\
& \times \text{tr}\left(\left(\sum_{j=1}^{m} \frac{\boldsymbol{x}_j \boldsymbol{x}_j^T}{(\sigma_{e,j}^2 + \sigma_{v,k}^2)}\right)^{-1} \left(\sum_{j=1}^{m} \frac{\boldsymbol{x}_j \boldsymbol{x}_j^T}{(\sigma_{e,j}^2 + \sigma_{v,k}^2)^2}\right)\right).
\end{aligned} \tag{4.57}$$

Further, (4.56) might be interpreted as a measure of the difference between the components. Replacing ξ_{ik} by an estimate, finally, the following approximation is obtained:

$$\widehat{\mathrm{MSE}}(\hat{\mu}_i^{\mathrm{mix}}|\hat{\mu}_i^{\mathrm{Dir}}) \approx \sum_{k=1}^{K} \hat{\xi}_{ik}\widehat{\mathrm{MSE}}_k + \sum_{k=1}^{K} \hat{\xi}_{ik}(\hat{\mu}_{ik}^{*\mathrm{FH}} - \hat{\mu}_i^{\mathrm{mix}})^2 \tag{4.58}$$

Chapter 5

Simulation Studies

5.1 Introduction

A common tool to evaluate the performance of a new estimator is to perform a simulation study. Generally, Monte Carlo (MC) simulation studies are the "process of sampling repeatedly from either a fixed population or a statistical model" (Zimmermann, 2015). In each run, selected results, e.g. point or variance estimates, are obtained from the estimator of interest and a selection of competing approaches. The distribution of results over the simulation runs (MC-distribution) can then be used to assess the performance of the estimator. Overall, this allows to evaluate estimation methods in a controlled environment, i.e. under specific well-chosen scenarios. This can be the only viable approach to assess the properties of an estimator if analytical solutions are not obtainable. But even if such solutions exist, simulation studies can result in surprising findings and shed light on peculiarities of an estimator that otherwise might have been missed (Alfons, Filzmoser, Hulliger, Kolb, Kraft, Münnich and Templ, 2011). Overall, simulations help to arrive at a informed judgement on the suitability of an estimator in a specific setting. With the rising availability of computational power they have increasingly become feasible and are a standard tool of studies in small area estimation. See Burgard, Münnich and Zimmermann (2016); Chandra and Chambers (2016); Molina and Rao (2010); Salvati, Chandra, Ranalli and Chambers (2010); Schmid, Tzavidis, Münnich and Chambers (2016) and Wagner, Münnich, Hill, Stoffels and Udelhoven (2017) for some examples.

Broadly two different frameworks for MC-simulations can be distinguished: model-based and design-based simulation studies. While in a model-based simulation

study, observations are generated from the model in each run, in a design-based study the samples are drawn from a fixed population. Randomisation is thus with respect to the design (See Zimmermann (2015), Münnich (2014), or Burgard (2013) for a more elaborate taxonomy, differentiating between four and six types of simulation studies, respectively).

Generally, model-based simulation studies are deemed the adequate approach to test the performance of new model-based estimators under different assumptions for the population. This, of course, implies both an assessment of the quality of the estimator when model assumptions are fulfilled and an evaluation of consequences of controlled violations of central assumptions (Chandra and Chambers, 2016; Zimmermann, 2015). They can be considered as a "check whether [a] new procedure really works" (Zimmermann, 2015).

Of course all estimators eventually are meant to be applied to real data. Then all models are only approximations of the reality. Additionally, practitioners are confronted with one single sample from a finite population, usually resulting from a complex survey process. This situation is mimicked with a design-based simulation, which therefore is a valuable device to judge the performance of an estimation strategy in a real-data application. It allows to evaluate the performance of a model-based estimator in a design-based framework, e.g. including assessments of the interplay of estimation method and sampling design. Moreover, a design-based simulation set-up implies that features of areas are held fixed (Salvati et al., 2010) so that no averaging over peculiarities veils the performance of the estimators for extreme cases. Results can vary significantly from those obtained in a model-based simulation (Burgard, Kolb, Merkle and Münnich, 2017) and reveal surprising features of the methods at hand (Alfons et al., 2011). If the aim is to evaluate the estimator in a close-to-reality scenario, the study however requires careful planning to build a set-up that is able to appropriately reproduce the complex interplay of characteristics of the population, the sampling design and data processing in a real survey (see Alfons et al. (2011); Münnich, Schürle, Bihler, Boonstra, Knottnerus, Nieuwenbroek, Haslinger, Laaksonen, Wiegert, Eckmair, Quatember, Wagner, Renfer and Oetliker (2003) and Burgard et al. (2017) as well as references therein for further discussions and sophisticated examples of synthetic datasets for design-based simulation studies).

In what follows, two model-based simulation studies are performed in order to arrive at a judgement of the general functionality and appropriateness of the suggested estimators. A more demanding design-based study in a complex close-to-reality setting to complement the findings is an important task for future research.

Common measures for the evaluation of results are the MC Relative Bias (RBIAS)

as well as the MC Relative Root Mean Square Error (RRMSE) (see Burgard, 2013; Zimmermann, 2015, for a more comprehensive discussion of quality measures for simulation studies in SAE).

For a simulation study with R MC-replications, the RBIAS of an estimator $\hat{\mu}$ is defined as

$$\text{RBIAS}_i := \frac{1}{R}\sum_{r=1}^{R}\frac{(\hat{\mu}_{i,r} - \mu_{i,r})}{\mu_{i,r}}. \tag{5.1}$$

$\hat{\mu}_{i,r}$ denotes the estimate obtained for area i in replication r and $\mu_{i,r}$ is the respective true value. RBIAS_i takes values in $(-\infty, \infty)$. Obviously, a result close to zero is desirable.

Further, the RRMSE is given by

$$\text{RRMSE}_i := \sqrt{\frac{1}{R}\sum_{r=1}^{R}\frac{(\hat{\mu}_{i,r} - \mu_{i,r})^2}{\mu_{i,r}^2}} \tag{5.2}$$

The RRMSE takes values in $(0, \infty)$. Again, an RRMSE close to 0 is the desired result.

Alternatively, the MC Bias (BIAS)

$$\text{BIAS}_i := \frac{1}{R}\sum_{r=1}^{R}(\hat{\mu}_{i,r} - \mu_{i,r}) \tag{5.3}$$

and Root Mean Square Error (RMSE)

$$\text{RMSE}_i := \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\mu}_{i,r} - \mu_{i,r})^2} \tag{5.4}$$

can be reported instead of the relative measures introduced above. This might be sensible if some of the true values are close to zero such that the considered measure takes a large value even for a small deviation between estimate and true value.

Additionally, to these area-specific measures, sometimes averages over all areas are considered. This allows to summarize results in an even compacter way and focusses on the overall performance in cases where compared estimators have

strengths or weaknesses for some specific areas. Relevant performance measures are the Mean Absolute Relative Bias (MARB)

$$\text{MARB} := \frac{1}{m}\sum_{i=1}^{m}|\text{RBIAS}_i| \tag{5.5}$$

and the Average Relative Root Mean Square Error (ARRMSE)

$$\text{ARRMSE} := \frac{1}{i}\sum_{i=1}^{m}\text{RRMSE}_i. \tag{5.6}$$

5.2 Area-level Simulation

5.2.1 Setting

The proposed estimators based on a mixture of area-level models with or without concomitant variables were tested in a model-based simulation study with 1000 runs. In each MC-iteration, the subgroup-label vector $\boldsymbol{z}_i$ was generated by drawing once from the categories 1 to K with probabilities $\lambda_1, \ldots, \lambda_K$ and the true mean $\mu_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k + v_{i,k}, i = 1, \ldots, 200$ was then generated from the respective component-model. The covariates $\boldsymbol{X}$ were held fixed over the simulation runs. An m-vector of design variances $\sigma^2_{e,i}$ was generated from $\sigma^2_{e,i} \sim unif(0.6, 0.24)$ in each run and the direct estimate was obtained as $\mu_i^{\text{Dir}} = \mu_i + e_i$ with $e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2_{e,i})$. Additionally, 50 areas where only the auxiliary information $\boldsymbol{x}_i$ and w_i are given were generated in order to also evaluate the performance of the suggested methods in case of non-sampled areas, i.e. areas for which no direct estimate is available. Finally, additional covariates $\boldsymbol{w}$ for the submodel for the mixture weights were generated in order to analyse the performance of the mixture-based estimator with concomitant variables.

Overall, four populations (Population 1 to 4) and two different settings with respect to w (Setting A and B) were considered: The populations were designed to represent different scenarios with respect to the clustered structure: While population 1 is a homogeneous population with $K = 1$, with population 2 and 3 two settings were considered where the areas actually are segmented into two equally sized subgroups. The component densities in populations 2 are clearly separated whereas the components in population 3 were designed as partly overlapping. With population 4 a scenario with unequally sized clusters was considered. Table 5.1 gives an overview.

Table 5.1: Populations in the simulation study

Population 1 homogeneous population	*Population 2* clearly separated components	*Population 3* partly overlapping components	*Population 4* unequally sized clusters
$K = 1$	$K = 2$		
$\lambda_k = 1/K$ for all k			$\lambda_1 = 0.15$, $\lambda_2 = 0.85$
$\boldsymbol{\beta}_{k=1} = (8.5, 0.2, 0.2)$	$\boldsymbol{\beta}_{k=1} = (9, 0.5, -0.25)$ $\boldsymbol{\beta}_{k=2} = (8.5, -0.5, 0.4)$	$\boldsymbol{\beta}_{k=1} = (11.5, 0.2, -0.1)$ $\boldsymbol{\beta}_{k=2} = (5, -0.2, 0.3)$	$\boldsymbol{\beta}_{k=1} = (9, 0.5, -0.25)$ $\boldsymbol{\beta}_{k=2} = (8.5, -0.5, 0.4)$
$\sigma_u^2 = 0.7$	$\sigma_{u,k}^2 = 0.7$ for all k		
$x_1 = \mathbf{1}$,$x_2 \sim N(-4, 2)$ and $x_3 \sim N(3, 2)$			
$\sigma_{e,i}^2 \sim unif(0.24, 0.6)$			

Additional to the four distinct scenarios, two different settings (Setting A and B) with respect to the concomitant variable $\boldsymbol{w}$ were considered. For the first setting (Setting A), $\boldsymbol{w}$ was designed to be highly correlated with the true cluster membership in the clustered populations 2 to 4. Further, the consequences of modelling the mixture weights through covariates without explanatory power of a given clustering structure were analysed (Setting B). For both cases an $m \times 2$ set of possible covariate values was constructed once. This was done by drawing an m-vector $\boldsymbol{w}_1$ from $\mathcal{SN}(-0.6, 0.275, 3)$ and an m-vector $\boldsymbol{w}_2$ from $\mathcal{SN}(0.6, 0.275, -3)$, where $\mathcal{SN}(\kappa, \omega, \rho)$ denotes the skew normal distribution with location κ, scale ω and skewness parameter ρ, respectively. For this purpose, `rsnorm` from the package `fGarch` was used.

For Setting A the values were assigned according to the respective component membership of the areas without error, i.e. for an area i in component k, w_i was set to be the ith value in $\boldsymbol{w}_k$. For Setting B, w_i was drawn randomly from the vector of K candidate values for area i. Figure 5.1 illustrate the resulting histograms of $\boldsymbol{w}$ for one exemplary run in a 2-component scenario.[1] This strategy of course is meaningless for the homogeneous population 1 where component membership is the same for all areas. Therefore, a slightly different specification of the two setting was chosen. In Setting A, an artificial clustering structure was imposed, i.e. each area was assigned to one of K clusters with probability λ_k. As in the clustered populations, w_i was then accordingly chosen from the vector of candidate values. This results in a bimodal distribution of $\boldsymbol{w}$ and thus implies the case of imposing information on a clustered pattern in case of a population that is homogeneous with respect to the main model. For Setting B, w_i was drawn from a uniform distribution $unif(-1, 1)$ to model a case where the concomitant variable does not introduce any additional clustering information.

[1] Note that while the $m \times K$-matrix of possible candidate values is fixed over MC-replications, the resulting vector of concomitant variables $\boldsymbol{w}$ is not, because component membership of areas is random.

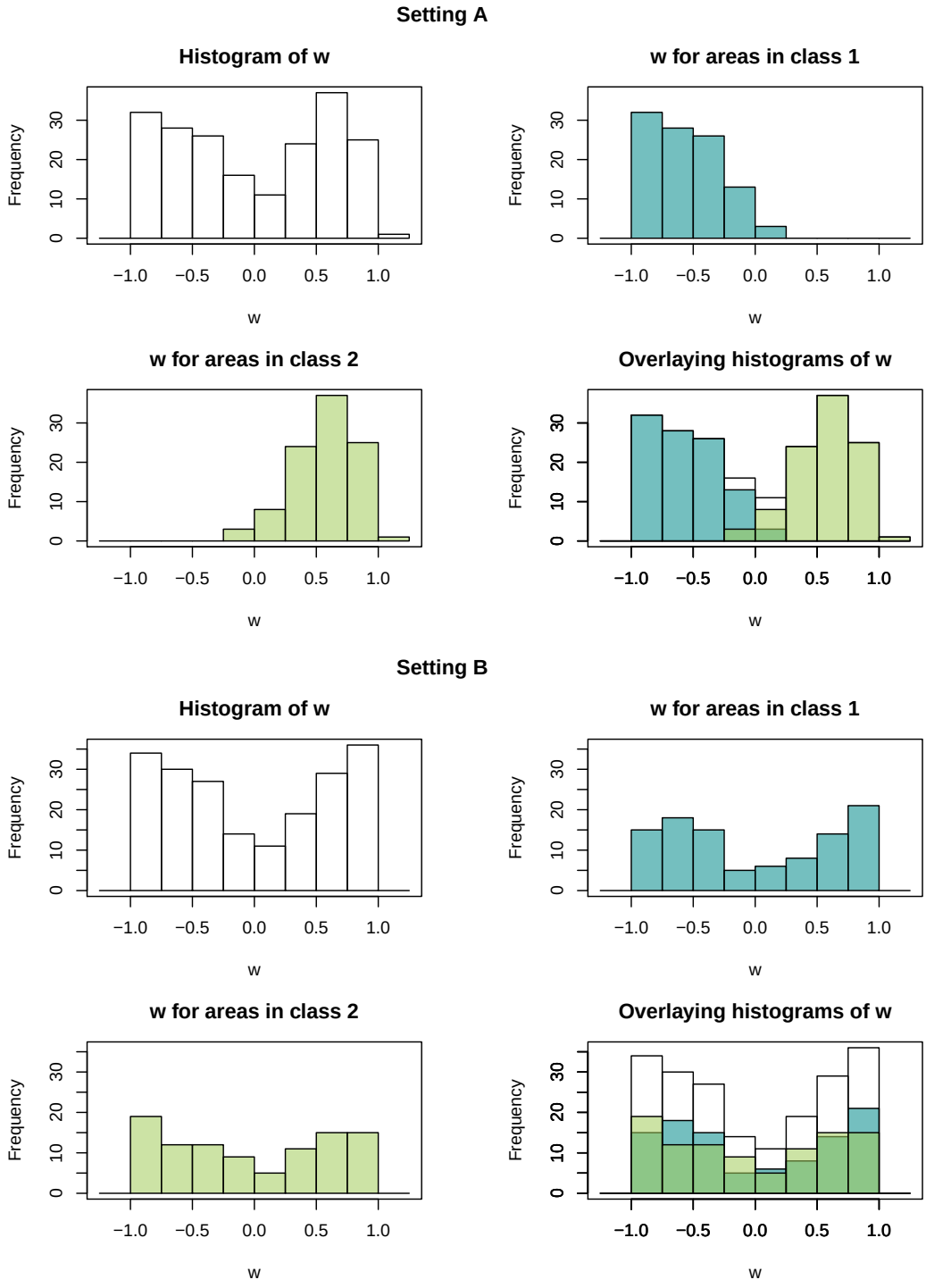

Figure 5.1: Histograms of $\boldsymbol{w}$ for exemplary MC-run ($K = 2$)

Overall, the simulation study is designed to reflect a scenario where the areas are clustered into K actually existing but latent subgroups with a true subgroup-specific functional relationship between μ_i and main-model covariates $\boldsymbol{x}$. Correspondingly, the direct estimates are generated according to the assumptions drawn for the component they belong to. There is a second set of covariates $\boldsymbol{w}$ available, which is employed to model the mixing proportions in the submodel in order to support segmentation and characterize the clusters. This submodel can be understood as an *a-posteriori* tool utilized to understand the data. It is not however some kind of true data-generating process in the sense that there is a true area-specific mixture weight λ explicitly defined through the submodel. This simulation scenario matches the intended purpose of the submodel in the context of small area estimation in the case of latent subgroups.

Note that by construction, $\boldsymbol{X}$ contains no systematic information about component membership. In practical applications it might often be reasonable to assume, that auxiliary information vary between different subgroups of areas. This can be expected to further support the identification of subgroups in the estimation process. Thus, the chosen setting of drawing the covariates from a single distribution, can be regarded as a conservative simulation approach.

Table 5.2: Estimators in the simulation study

DIR	Direct estimator
FH	Standard FH-estimator
FHmix	Mixture-based FH estimator without concomitant variables
FHmixconc	Mixture-based FH with concomitant variables
clustFH.k	K-means clustering-based FH estimator. Obtained by (1) k-means clustering of observations based on $\boldsymbol{w}$ and (2) subsequent estimation of cluster-specific FH-models
clustFH.mix	Cluster-based FH estimator. Obtained by (1) clustering of observations based on $\xi_{i,k}$ and (2) subsequent estimation of cluster-specific FH-models

An overview of estimators considered in the study is given in Table 5.2 and Table 5.2. Table 5.2 lists the estimating approaches considered for the sampled areas. Additional to the proposed mixture-based area-level estimators (FHmix)

and (FHmixconc), the standard FH-estimator (FH) is included in the study as a benchmark. Further, with clustFH.k and clustFH.mix two alternative cluster-based approaches are considered. Parameter estimation for the mixture-based approaches was performed applying Version 1 of the EM algorithm described in Section 4.3 and 4.5.

In Section 4.4, a concomitant variable mixture model-based estimator for out-of-sample prediction was discussed as a possible approach for unsampled areas. Its performance for out-of-sample prediction is evaluated against reference approaches based on the standard FH-model, FHmix and clustFH.k. Respective details can be taken from Table 5.3.

Table 5.3: Estimation strategies for unsampled areas

FH.oos	Synthetic estimator
FHmix.oos	Weighted mean of synthetic estimators, where the weights are the estimated (fixed) mixing-proportions λ_k
FHmixconc.oos	Weighted mean of synthetic estimators, where weights are given by the area-specific out-of-sample predicts from the sub-model
clustFH.k.oos	clustFH.k as cluster-specific synthetic estimator, where unsampled areas are assigned to one of K clusters by k-means clustering based on $\boldsymbol{w}$.

Additionally, the performance of two different criteria for estimating K, namely the ICL-BIC and the BIC, was evaluated. For this purpose, each criterion was obtained for FHmix and FHmixconc in Setting A and B letting K grow from 1 to 5 and $\widehat{K}$ was chosen as the specification minimizing the respective criterion (see Section 3.6).

5.2.2 Results

A first step in each application of mixture models is assessing the number of components in the population. Table 5.4 evaluates the performance of the suggested model selection criteria ICL-BIC and BIC in this regard. In each simulation run, K, was estimated by calculating the BIC and the ICL-BIC letting K grow from 1 to 4 (compare Section 3.6). Table 5.4 presents the percentage of simulation runs

where the respective estimate $\widehat{K}$ for K was obtained. First considering results for FHmix, it can be seen that for the homogeneous population 1 as well as for the clearly clustered populations 2 and 4 both criteria obtain reasonably good results. The performance of the BIC in detecting present clusters is, however, notably better. For population 3 the results of the criteria differ in a way consistent with the properties of ICL-BIC and BIC described in Section 3.6. The ICL-BIC, which penalizes poorly separated components and is thus the appropriate criterion in a clustering context, results in $\widehat{K} = 1$. The BIC, which is generally considered to be a suitable measure if the aim is density estimation for a heterogeneous population, estimates the true number of components, i.e. $K = 2$, at least in 78.6 percent of simulation runs. As discussed in Section 3.6, the foremost motivation of applying mixtures in SAE is to find a model that suits the data well in order to predict the statistic of interest in a heterogeneous population. Identifying distinct clusters is of secondary interest. Thus, it seems recommendable to follow the result of the BIC. It can, nevertheless be insightful to consider the ICL-BIC as well: A conflicting result of the two measures might hint at a scenario of poorly separated, but nevertheless existent, clusters.

Results for FHmixconc in Setting A, show that including further information on a clustered population in the estimation process heightens the probability of correctly estimating K in case of a clustered population. This is especially true for the partly overlapping components in population 3 where now both criteria almost certainly correctly result in $\widehat{K} = 2$. The performance of the ICL-BIC in population 2 and 4 is also notably improved. Setting B reveals the consequences of imposing a "false clustering structure". Both criteria still perform almost equally well as in an approach without concomitant variables. Thus, while the inclusion of suitable information does improve the estimation performance, the inclusion of an additional "wrong" clustered pattern through the submodel does not seem to "mislead" the model selection criteria into assuming a clustered structure for the main model, too.

Table 5.4: Simulation results for estimating K

Population	*Criterion*	*Result (Percentage of 1000 MC-runs)*											
		FHmix				FHmixconc Setting A				FHmixconc Setting B			
		$\widehat{K}=1$	$\widehat{K}=2$	$\widehat{K}=3$	$\widehat{K}=4$	$\widehat{K}=1$	$\widehat{K}=2$	$\widehat{K}=3$	$\widehat{K}=4$	$\widehat{K}=1$	$\widehat{K}=2$	$\widehat{K}=3$	$\widehat{K}=4$
Population 1	ICL-BIC	**100**	0	0	0	**100**	0	0	0	**100**	0	0	0
	BIC	**100**	0	0	0	**100**	0	0	0	**100**	0	0	0
Population 2	ICL-BIC	6.5	**93.5**	0	0	0	**100**	0	0	5.6	**94.4**	0	0
	BIC	0	**100**	0	0	0	**99.9**	0.1	0	0	**99.9**	0.1	0
Population 3	ICL-BIC	100	**0**	0	0	0	**99.6**	0.4	0	100	**0**	0	0.1
	BIC	21.4	**78.6**	0	0	0	**99.9**	0.1	0	20.6	**79.4**	0	0
Population 4	ICL-BIC	9.1	**90.9**	0	0	0	**100**	0	0	9.1	**90.9**	0.4	0
	BIC	0	**99.5**	0.5	0	0	**100**	0	0	0	**99.1**	0.9	0

Table 5.5 and 5.6 present summarizing statistics for the estimated model parameters of the main model in Setting A and Setting B, respectively. Listed are the mean and standard deviation of estimation results over the simulation runs. Note that, due to the label-switching problem described in Section 3.4, a meaningful representation of component- or cluster-specific parameter estimates for the mixture- and cluster-based estimators requires ordering and subsequent relabelling of estimation results for the two components. This was achieved by imposing the restriction $\hat{\beta}_{1,k=1} > \hat{\beta}_{1,k=2}$, a strategy that has proven to serve well in the present context.

It can be taken from these tables, that the estimation results obtained for the mixture based estimators in the clustered populations are generally quite accurate. As to be expected, results are best for the clearly clustered and balanced population 2 and for the larger cluster ($k = 2$) in population 4, but the performance for population 2 and component 1 in population 4 is also good. In the homogeneous population 1 the mixture-based estimators seem to artificially reduce the true model variance be separating observations into two homogeneous subgroups. Further, regarding results for FHmixconc it can be seen that the inclusion of suitable concomitant variables (Setting A in Table 5.5) tends to support the separation of components and to stabilize the estimation of model parameters. This is most obvious in population 3 and the unbalanced population 4, where results for the smaller component 1 are improved. Considering results for FHmixconc obtained in Setting B, it can be seen that the inclusion of uncorrelated covariates in the submodel does not negatively affect results. FHmixconc performs equally well as FHmix.

Considering the benchmark approach of FH, it can be seen that it performs well in the homogenous population 1 but, as to be expected, results in a large estimated model variance in all clustered populations.

Table 5.5: Estimated parameters (Setting A): Mean and standard deviation over simulation runs

	$k=1$								$k=2$							
	Population 1															
	$\hat{\sigma_v}^2$		Icept		β_1		β_2		$\hat{\sigma_v}^2$		Icept		β_1		β_2	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
FH	0.67	0.21	8.52	0.28	0.20	0.05	0.20	0.06								
FHmix	0.48	0.45	8.97	0.77	0.31	0.11	0.19	0.17	0.45	0.45	8.07	0.74	0.10	0.10	0.20	0.16
FHmixconc	0.46	0.47	9.00	0.76	0.32	0.11	0.19	0.16	0.42	0.45	8.03	0.78	0.09	0.10	0.20	0.17
FHclust.k	0.65	0.28	8.69	0.37	0.24	0.06	0.20	0.08	0.64	0.29	8.33	0.37	0.16	0.06	0.20	0.08
FHclust.mix	0.30	0.70	9.29	1.29	0.38	0.18	0.21	0.27	0.27	0.71	7.73	1.34	0.02	0.18	0.19	0.28
	Population 2															
	$\hat{\sigma_v}^2$		Icept		β_1		β_2		$\hat{\sigma_v}^2$		Icept		β_1		β_2	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
FH	9.64	0.68	8.77	0.61	0.01	0.12	0.07	0.14								
FHmix	0.64	0.35	9.02	0.52	0.50	0.08	-0.25	0.09	0.62	0.35	8.52	0.51	-0.50	0.09	0.40	0.09
FHmixconc	0.63	0.29	8.99	0.43	0.49	0.08	-0.25	0.08	0.62	0.29	8.54	0.42	-0.50	0.07	0.39	0.08
FHclust.k	1.74	0.78	8.93	0.53	0.47	0.08	-0.21	0.11	1.89	0.70	8.43	0.43	-0.45	0.08	0.42	0.09
FHclust.mix	0.43	0.29	8.78	0.61	0.47	0.10	-0.24	0.09	0.42	0.29	8.76	0.59	-0.47	0.10	0.38	0.09
	Population 3															
	$\hat{\sigma_v}^2$		Icept		β_1		β_2		$\hat{\sigma_v}^2$		Icept		β_1		β_2	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
FH	4.21	0.46	8.24	0.55	-0.00	0.09	0.10	0.09								
FHmix	0.71	0.58	11.56	0.60	0.21	0.10	-0.11	0.11	0.70	0.54	4.96	0.59	-0.22	0.10	0.31	0.12
FHmixconc	0.57	0.28	11.51	0.42	0.20	0.08	-0.09	0.08	0.55	0.28	5.03	0.43	-0.19	0.07	0.30	0.08
FHclust.k	1.07	0.43	11.26	0.46	0.19	0.08	-0.07	0.08	1.21	0.43	5.31	0.45	-0.19	0.08	0.27	0.08
FHclust.mix	0.26	0.34	11.46	0.65	0.18	0.11	-0.07	0.12	0.26	0.31	5.06	0.62	-0.18	0.10	0.27	0.13
	Population 4															
	$\hat{\sigma_v}^2$		Icept		β_1		β_2		$\hat{\sigma_v}^2$		Icept		β_1		β_2	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
FH	5.25	0.92	8.61	0.48	-0.35	0.09	0.30	0.11								
FHmix	0.87	1.46	9.30	1.33	0.52	0.23	-0.26	0.23	0.62	0.37	8.50	0.37	-0.50	0.06	0.40	0.08
FHmixconc	0.60	0.54	9.10	0.97	0.50	0.16	-0.27	0.17	0.66	0.23	8.53	0.31	-0.50	0.06	0.39	0.06
FHclust.k	6.46	2.18	8.73	1.12	0.28	0.24	-0.03	0.25	0.90	0.72	8.49	0.36	-0.50	0.06	0.39	0.08
FHclust.mix	0.56	1.31	8.43	1.37	0.42	0.22	-0.20	0.22	0.57	0.36	8.48	0.37	-0.51	0.06	0.40	0.08

Table 5.6: Estimated parameters (Setting B): Mean and standard deviation over simulation runs

	$k=1$								$k=2$							
	Population 1															
	$\hat{\sigma_v}^2$		Icept		β_1		β_2		$\hat{\sigma_v}^2$		Icept		β_1		β_2	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
FH	0.67	0.21	8.52	0.28	0.20	0.05	0.20	0.06								
FHmix	0.48	0.45	8.97	0.77	0.31	0.11	0.19	0.17	0.45	0.45	8.07	0.74	0.10	0.10	0.20	0.16
FHmixconc	0.46	0.46	9.00	0.78	0.31	0.11	0.19	0.17	0.43	0.46	8.04	0.78	0.09	0.10	0.20	0.17
FHclust.k	0.64	0.28	8.69	0.37	0.24	0.06	0.19	0.08	0.65	0.29	8.35	0.37	0.16	0.06	0.20	0.08
FHclust.mix	0.30	0.70	9.29	1.29	0.38	0.18	0.21	0.27	0.27	0.71	7.73	1.34	0.02	0.18	0.19	0.28
	Population 2															
	$\hat{\sigma_v}^2$		Icept		β_1		β_2		$\hat{\sigma_v}^2$		Icept		β_1		β_2	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
FH	9.64	0.68	8.77	0.61	0.00	0.12	0.07	0.15								
FHmix	0.63	0.35	9.02	0.52	0.50	0.09	-0.25	0.09	0.62	0.35	8.52	0.51	-0.50	0.09	0.40	0.09
FHmixconc	0.63	0.35	9.02	0.52	0.50	0.09	-0.25	0.09	0.62	0.35	8.52	0.51	-0.50	0.09	0.39	0.09
FHclust.k	9.42	1.05	9.12	0.79	0.10	0.14	0.06	0.20	9.49	1.07	8.43	0.81	-0.09	0.15	0.09	0.21
FHclust.mix	0.43	0.29	8.78	0.61	0.47	0.10	-0.24	0.09	0.42	0.29	8.76	0.60	-0.47	0.10	0.38	0.09
	Population 3															
	$\hat{\sigma_v}^2$		Icept		β_1		β_2		$\hat{\sigma_v}^2$		Icept		β_1		β_2	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
FH	4.21	0.46	8.24	0.55	-0.00	0.09	0.10	0.09								
FHmix	0.71	0.59	11.56	0.61	0.22	0.10	-0.11	0.11	0.71	0.57	4.97	0.60	-0.22	0.10	0.31	0.11
FHmixconc	0.73	0.62	11.55	0.63	0.22	0.11	-0.11	0.12	0.72	0.61	4.97	0.62	-0.22	0.10	0.31	0.12
FHclust.k	4.11	0.66	8.60	0.72	0.07	0.11	0.09	0.13	4.14	0.67	7.90	0.70	-0.07	0.11	0.11	0.13
FHclust.mix	0.26	0.36	11.47	0.62	0.18	0.11	-0.07	0.12	0.27	0.33	5.06	0.61	-0.18	0.10	0.27	0.12
	Population 4															
	$\hat{\sigma_v}^2$		Icept		β_1		β_2		$\hat{\sigma_v}^2$		Icept		β_1		β_2	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
FH	5.29	0.92	8.63	0.48	-0.34	0.09	0.30	0.11								
FHmix	0.86	1.42	9.29	1.33	0.52	0.23	-0.26	0.23	0.61	0.28	8.51	0.37	-0.50	0.06	0.40	0.08
FHmixconc	0.90	1.49	9.31	1.32	0.52	0.23	-0.26	0.23	0.61	0.29	8.51	0.36	-0.50	0.06	0.40	0.08
FHclust.k	5.57	1.25	8.90	0.64	-0.27	0.11	0.28	0.16	4.77	1.27	8.36	0.61	-0.42	0.11	0.31	0.15
FHclust.mix	0.55	1.24	8.44	1.37	0.42	0.22	-0.20	0.22	0.56	0.25	8.49	0.36	-0.50	0.06	0.40	0.08

The RRMSE and RBIAS of point estimates are illustrated in 5.2 and 5.3.

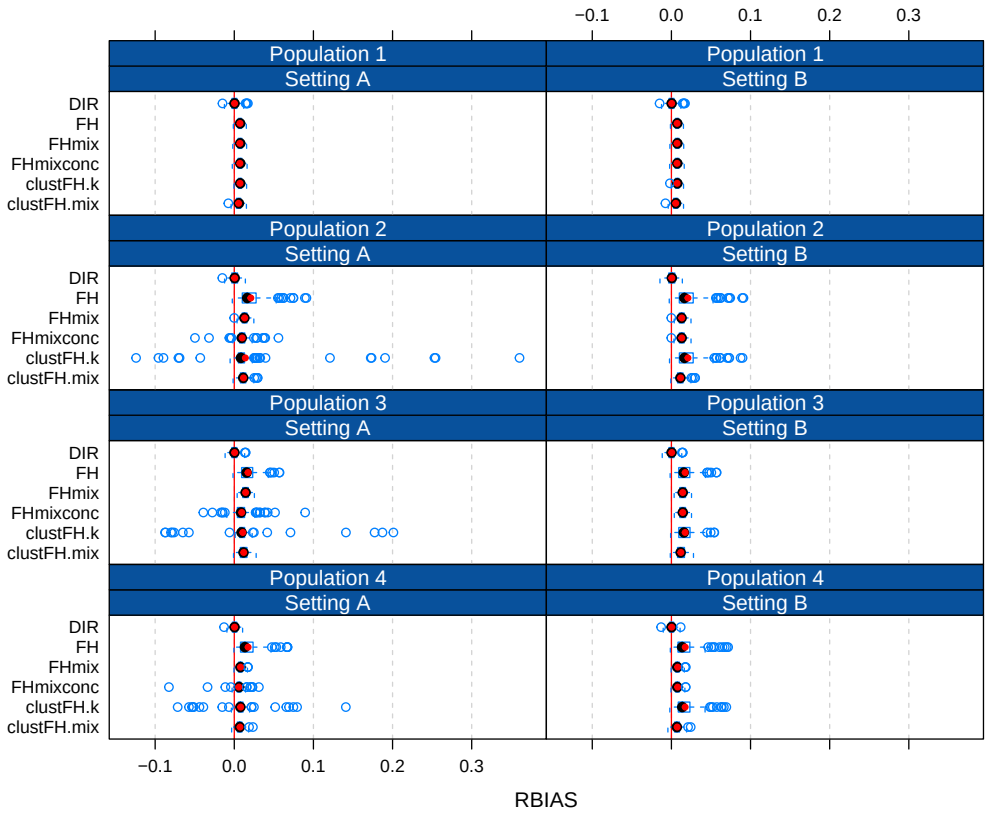

Figure 5.2: RBIAS

From the depiction of the RBIAS in Figure 5.2, it can be taken that application of small area techniques introduces bias compared to the design based direct estimator. This is an expected result and is, generally, accepted as a price to be paid for the reduction in variance. The RRMSE in Figure 5.3 reveals the performance of the considered estimators regarding these conflicting objects. Overall, the employment of small area methods reduces the median and mean RRMSE. An exemption is Population 4, where the improvement realized for some areas, which is expressed in a reduction of the median of the distribution, is outweighted by a large RRMSE for some outlying areas. This said, it obviously is of larger interest how the employment of the mixture-based estimators FHmix and FHmixconc compares to the application of the standard FH estimator and competing cluster-based approaches.

First of all, results for population 1 can be used to analyse the consequences of

incorrectly assuming a mixture distribution when the population actually is homogeneous. A crucial result was obtained: Even with this kind of misspecification the reduction of RRMSE, realized through the application of small area techniques instead of the employment of the direct estimator, is largely retained (see Figure 5.3). FHmix and FHmixconc perform almost equally well as FH. Even more importantly, Figure 5.2 shows that the employment of a mixture-based estimator in a homogeneous setting does not cause additional bias, i.e the mixture-based estimators FHmix and FHmixconc perform similar to the standard FH.

Regarding the clustered populations 2 to 4, the standard FH estimator tends to slightly overestimate the true value. Overall this bias is quite modest for the majority of areas, but there are few outlying areas with larger RBIAS. Further, there is almost no reduction of RRMSE compared to the direct estimator. This can be explained by the large estimated model variance (compare (Table 5.5 and 5.6) that leads to large shrinkage factors and, thus, a strong reliance on the design-based part of the small area estimator. Results can be improved when applying FHmix. This is especially true for the clearly clustered populations 2 and 4 and most strikingly visible in the upper tail of the distribution. Regarding the RRMSE, it can be seen that the introduction of mixture-based small area methods instead of the standard FH also generally reduces the RRMSE. As expected the realized improvement is larger in the clearly clustered populations 2 and 4 and less pronounced in population 3.

Regarding the inclusion of concomitant variables into the framework it can first be taken from Setting B that the inclusion of "wrong" covariates in the submodel does not negatively affect the prediction performance: FHmix and FHmixconc perform equally well in all scenarios, both in terms of RBIAS and RRMSE. If suitable concomitant variables are introduced into the framework (Setting A), a further reduction of the ARRMSE and MARB can be realized in all clustered populations. This is especially true for the overlapping components of mixpop 3. It however comes with the price of a larger RBIAS and RRMSE for some outlying areas.

While clustFH.k in Setting A also slightly reduces the MARB compared to FHmix, it results in a unacceptably large RBIAS for some areas. This is especially true in population 2. Regarding the RRMSE, it yields better results than FHmix for most areas, but again there are few areas for which the performance seems unacceptable. Further, clustFH.k never performs better than FHmixconc. In Setting B, clustFH.k performs similar to FH in all populations while FHmixconc still maintains the improvement realized through accounting for the clustered structure disregarding of the "false" clustering structure imposed through the submodel. Overall clustFH.k is, thus, clearly outperformed by the suggested mixture-based

estimators. clustFH.mix yields similar results as FHmix in the clearly clustered populations 2 and 4 but performs worse in population 3 and more pronounced in population 1.

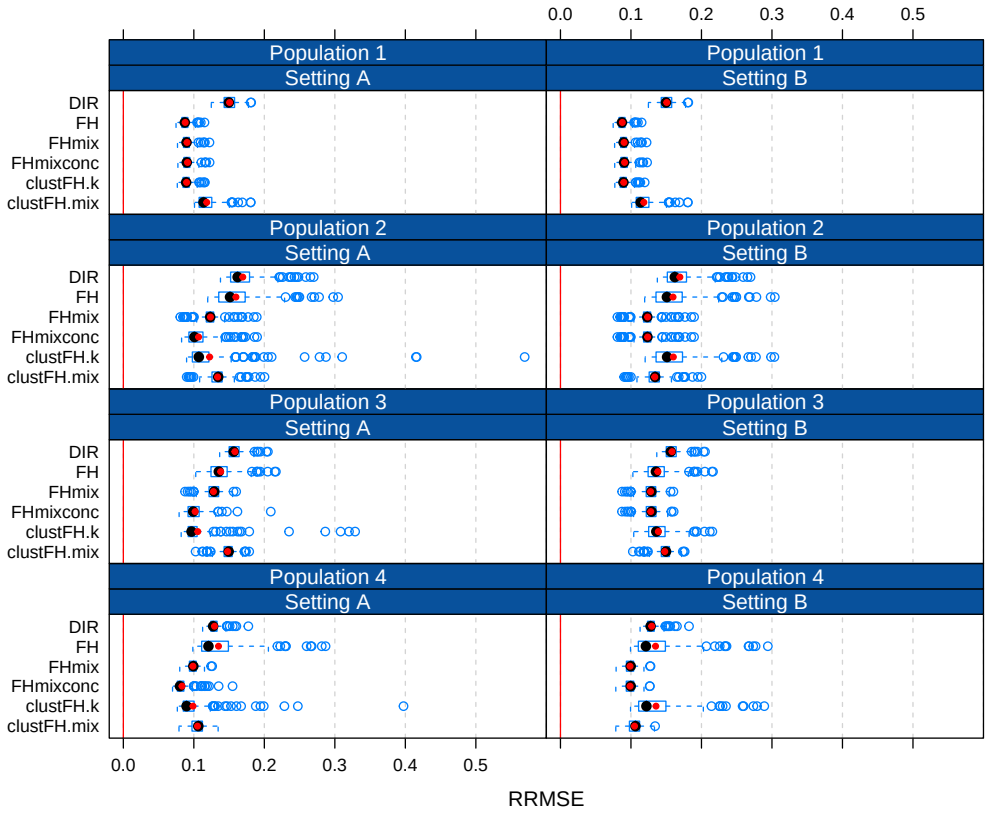

Figure 5.3: RRMSE

Note that, according to the standard model assumptions for finite mixture models, class membership was treated as random, i.e. $\boldsymbol{z}_i$ was drawn in each run. Summarizing statistics as the RBIAS_i and the RRMSE_i for i, thus, average over the component membership, veiling possible patterns in the estimation errors. An estimator that is unbiased in this marginal representation, is not necessarily unbiased when conditioning on $\boldsymbol{z}$. As in a real-data application, researchers are confronted with one single realisation, respective patterns are of great interest. Therefore, the analysis of simulation results is complemented by $\text{BIAS}_{i,k}$, which is defined as follows:

$$\text{BIAS}_{i,k} := \text{BIAS}_i|(z_{ik} = 1) := \frac{1}{\#(\boldsymbol{r}_k^*)} \sum_{r \in \boldsymbol{r}_k^*} (\hat{\mu}_{i,r} - \mu_{i,r}), \tag{5.7}$$

where $\boldsymbol{r}_k^*$ is a subset of the index of simulation runs containing all repetitions where i is generated from component model k such that $z_{ik} = 1$, and $\#(\boldsymbol{r}_k^*)$ is the cardinality of $\boldsymbol{r}_k^*$. Note that $E(\#(\boldsymbol{r}_k^*)) = \lambda_k \times R$, i.e. for populations 2 and 3 this measure on average is only based on 500 simulation runs. The BIAS is considered instead of the RBIAS because the differences in $\boldsymbol{x}_i^T \boldsymbol{\beta}_k$ otherwise complicate the comparison of component-specific results.

Figure 5.4 illustrates the results. For each estimator now two subgroup-specific results are depicted: FHmix1, for example, denotes the respective result for cluster 1, i.e. the $\text{BIAS}_{i,1}$ for the estimator FHmix. First regarding results for the clustered populations 2 and 3, it can be seen that there indeed is a subgroup-specific bias: In population 2, for example, all model-based clearly tend to overestimate the target statistic in subgroup 1 and to underestimate it in subgroup 2. FH performs worst. Results can be clearly improved if the clustered structure is appropriately accounted for. This is true for FHmix and – in Setting A – most strikingly for FHmxconc, that is able to reduce the mean absolute bias to almost zero. clustFH.k also yields good results, if suitable covariates $\boldsymbol{w}$ are available, but, obviously, fails to improve estimation results if no fitting covariates are available (Setting B).

Note that there is no true clustered structure in population 1. Thus, in conditioning on the subgroup-membership, here the "false" clustering structure imposed trough the concomitant variable $\boldsymbol{w}$ in Setting A was used. Again, this was meant to reveal possible hazardous consequences of including this additional information into the estimation framework. There are, however, no noticeable patterns in the results for FHmixconc and clustFH.k, the two estimators that use $\boldsymbol{w}$.

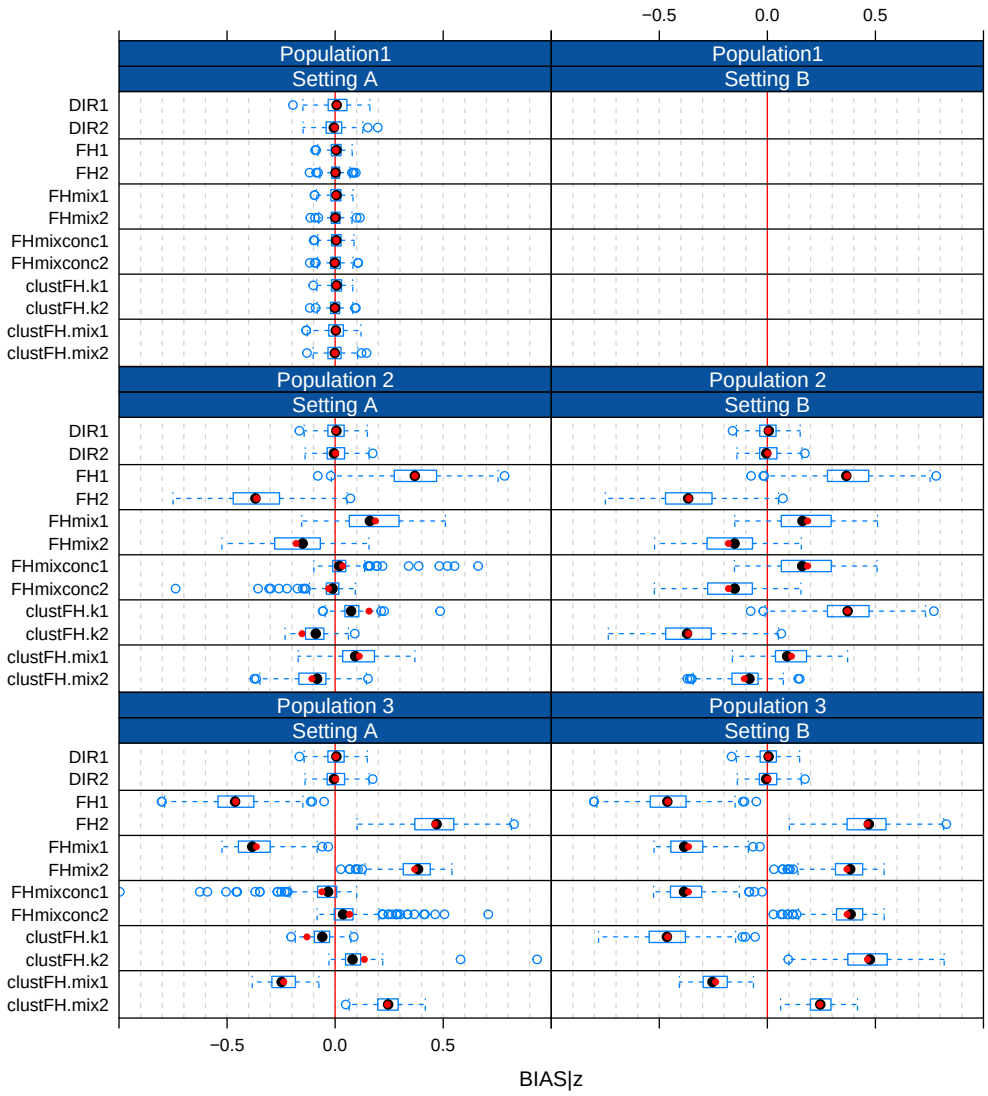

Figure 5.4: BIAS conditional on $\boldsymbol{z}$

Table 5.7 evaluates the clustering performance of the proposed method. It shows the average number of correctly assigned areas over the simulation runs (mean and standard deviation over 1000 runs). It can be seen that FHmixconc and FHmix perform almost equally well in the case of an uncorrelated auxiliary variable for the submodel (Setting B). As to be expected, they clearly outperform clustFH.k, where clustering is solely based on $\boldsymbol{w}$. When strong covariates are available (Setting A), FHmixconc yields much better results than FHmix and results in almost perfect clustering even for the overlapping clusters of population 3. Thus, cluster-

ing results can be improved considerably be introducing concomitant variables into the framework. At the same time, the approach again proves to be quite robust against misspecification in the sense that the inclusion of unfitting covariates in the submodel does not deteriorate clustering results.

Table 5.7: Number of areas correctly assigned to clusters ($m = 200$): Mean and standard deviation over simulation runs

		Setting A		*Setting B*	
		mean	std.dev	mean	std.dev
Population 2	FHmixconc	197.9	0.22	185.5	3.55
	FHmix	185.5	3.51	185.5	3.53
	clustFH.k	193.0	1.86	101.9	6.64
Population 3	FHmixconc	195.9	2.05	171.9	9.48
	FHmix	172.6	10.21	172.3	8.77
	clustFH.k	193.9	1.85	100.82	6.87
Population 4	FHmixconc	198.4	1.15	191.6	9.12
	FHmix	191.5	10.21	191.7	9.23
	clustFH.k	189.5	14.97	101.4	6.82

Results for out-of-sample prediction are illustrated in Figure 5.5 and 5.6. Again an important result is obtained from population 1: FHmix and FHmixconc perform as good as the standard small area method FH both in terms of RBIAS and of RRMSE. If, on the other hand, the population actually is heterogeneous, out-of-sample prediction using FH and FHmix results in a considerable bias and a large RRMSE. Given suitable information, the prediction result in this case can be improved considerably by employing FHmixconc or clustFH.k.

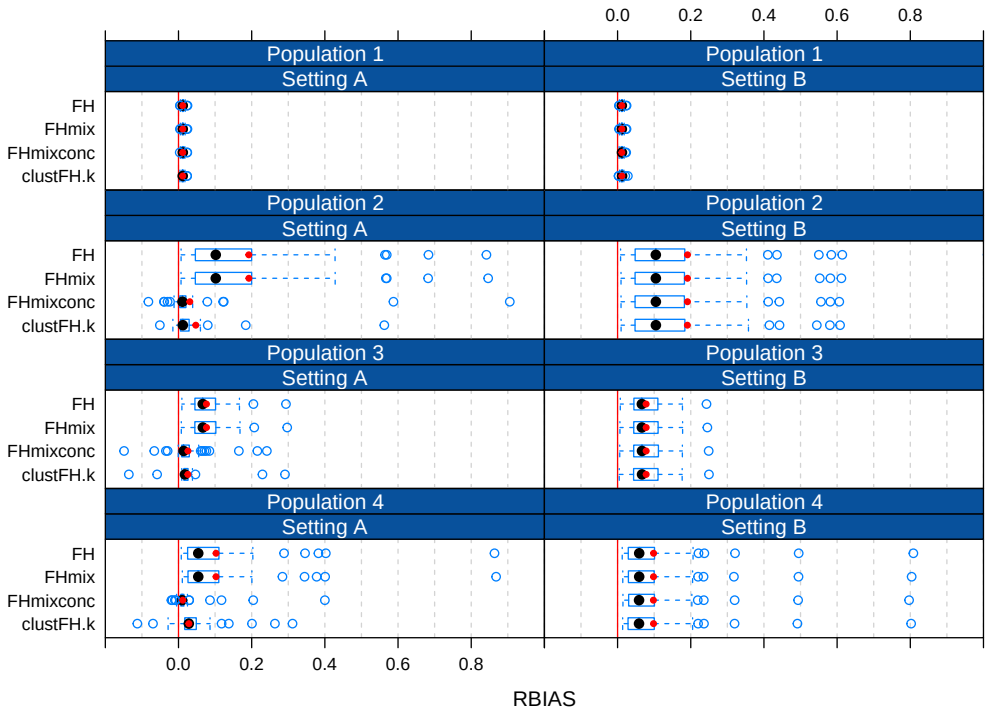

Figure 5.5: RBIAS for out-of-sample prediction

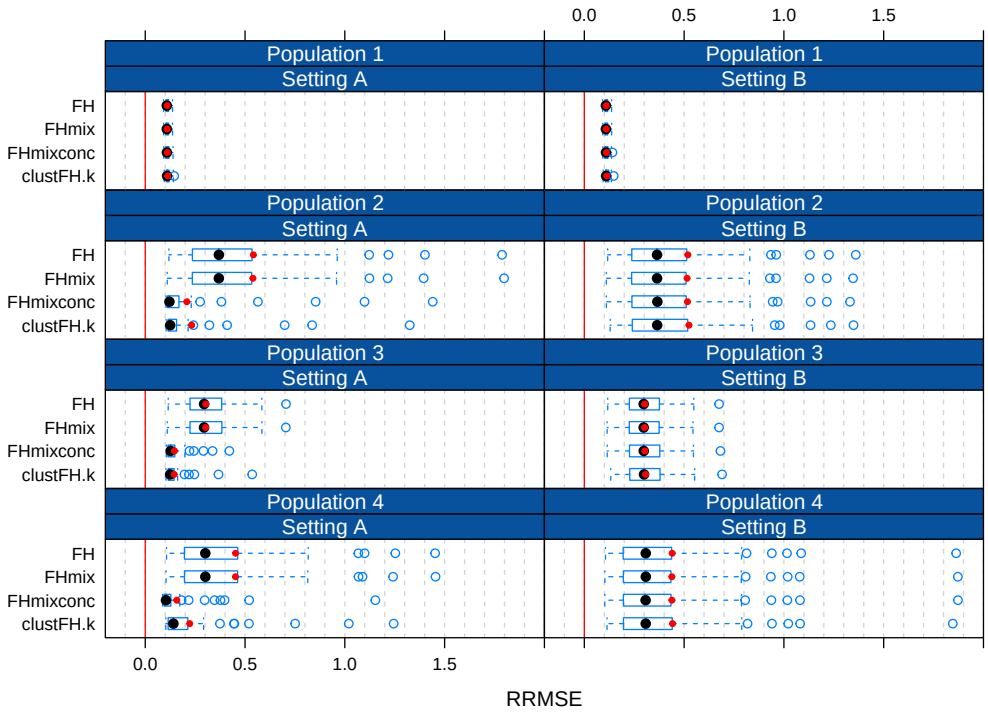

Figure 5.6: RRMSE for out-of-sample prediction

Finally, results for the suggested MSE estimator are evaluated. Figure 5.7 illustrates the BIAS of MSE estimation for the standard FH estimator and the mixture-based estimators FHmix and FHmixconc. The plot reveals that the suggested MSE approximation for FHmix and FHmixconc underestimates the true MSE. To put the observed bias in relation, Table 5.8 gives an overview of the average MSE, $\overline{\text{MSE}}_i := \frac{1}{R}\sum_{r=1}^{R}(\hat{\mu}_{i,r} - \mu_{i,r})^2$. Listed are the mean and standard deviation of $\overline{\text{MSE}}_i$ over the areas. In population 2 and setting A for example, a bias of around 0.05 roughly corresponds to 9% of the mean average MSE. While this may be considered as a good starting point, there is still room for improvement and further research seems necessary. It should most importantly include the assessment of the uncertainty introduced through the estimation of the posterior probabilities of component membership.

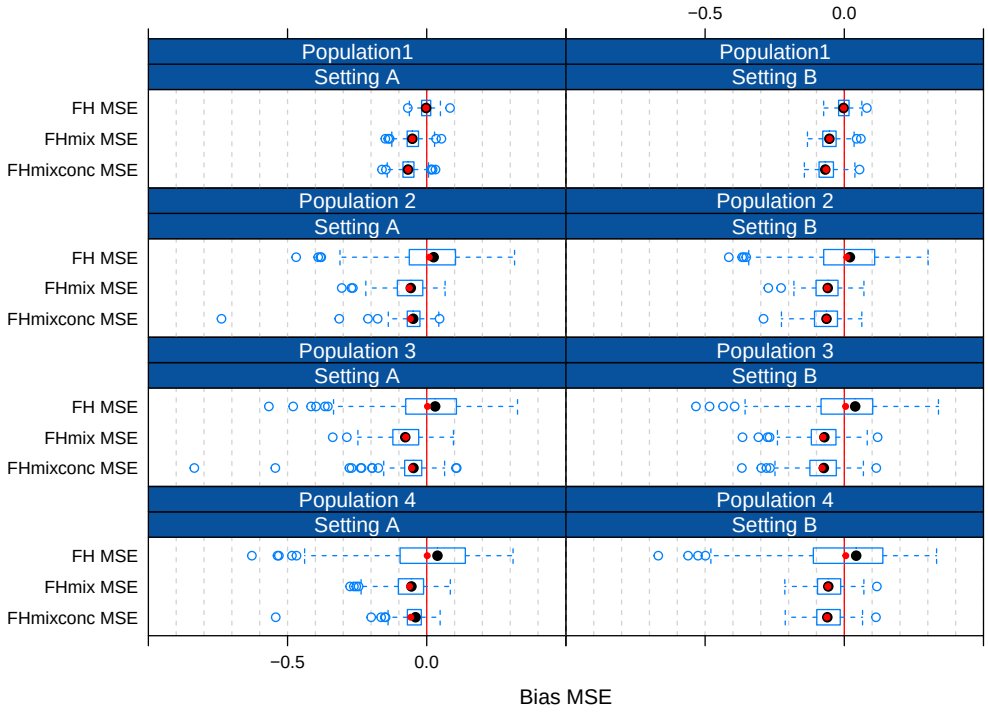

Figure 5.7: BIAS MSE estimator

		Setting A		*Setting B*	
		mean	std.dev	mean	std.dev
Population 1	FH MSE	0.488	0.027	0.488	0.030
	FHmix MSE	0.527	0.042	0.526	0.045
	FHmixconc MSE	0.519	0.040	0.520	0.043
Population 2	FH MSE	1.281	0.140	1.282	0.137
	FHmix MSE	0.566	0.158	0.800	0.154
	FHmixconc MSE	0.798	0.155	0.799	0.153
Population 3	FH MSE	1.094	0.152	1.093	0.152
	FHmix MSE	0.595	0.143	0.961	0.098
	FHmixconc MSE	0.958	0.097	0.961	0.097
Population 4	FH MSE	1.151	0.178	1.146	0.181
	FHmix MSE	0.552	0.186	0.706	0.101
	FHmixconc MSE	0.707	0.105	0.704	0.100

Table 5.8: Mean and standard deviation of average MSE

5.3 Unit-level Simulation

5.3.1 Setting

For evaluating the performance of the mixture of unit-level models under different scenarios, a finite population model-based simulation study (Zimmermann, 2015) with 1.000 MC-runs was performed. In each run, a population of 30,000 units, nested in 100 areas, was drawn from different population models. The nesting structure of units in areas was constant throughout the simulation, which also implies fixed area sizes N_i over the runs. They ranged from $N_i = 255$ to $N_i = 343$ with average area size of 300. Further, the covariates $\boldsymbol{X}$ were held fixed over all runs. As in the area-level study, additional covariates $\boldsymbol{w}$ were generated as submodel for the mixture weights. The corresponding $K \times m$ matrix was generated once. In each run, one sample of size $n = 600$ was drawn by simple random sampling, imposing the constriction that $n_i \leq 2$ for all areas $i = 1, \ldots, m$.

Seven different population models were considered to study the performance of the suggested approach under different scenarios (Population 1 to 7). Table 5.9 provides an overview. As in the study in the previous Section, populations were designed to represent different scenarios with respect to the existence of subgroups, that might possibly be encountered in real-data applications: First, with population 1 a homogeneous population with $K = 1$ is considered. Populations 2 and 3 are constructed to reflect a setting where the areas are segmented into two equal sized subgroups. The clusters in populations 2 are clearly separated whereas the subgroups in population 3, were designed as partly overlapping. Further, with population 4 and 5 two scenarios are considered where the subgroups are of unequal size. Population 6 is a highly clustered population with four equally sized subgroups of areas. In population 7, the covariates in the main model have no explanatory power in one of the two components. See Appendix A.2 for exemplary illustrations of the data.

Additional to the seven population models, two different settings (Setting A and B) with respect to the concomitant variable $\boldsymbol{w}$ were considered. For the first setting (Setting A) $\boldsymbol{w}$ was designed to be highly correlated with the true cluster membership. Further, the consequences of modelling the mixture weights through covariates without explanatory power of cluster membership were analysed (Setting B). For both cases an $m \times K$ set of possible covariate values was constructed once. For populations with $K = 2$ this was done by drawing an m-vector $\boldsymbol{w}_1$ from $\mathcal{SN}(-0.6, 0.275, 3)$ and an m-vector $\boldsymbol{w}_2$ from $\mathcal{SN}(0.6, 0.275, -3)$, where $\mathcal{SN}(\kappa, \omega, \rho)$ denotes the skew normal distribution with location κ , scale ω and skewness parameter ρ, respectively. For this purpose, `rsnorm` from the

package `fGarch` was used. For population 6 with $K = 4$, $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3$ and $\boldsymbol{w}_4$ were drawn from $\mathcal{N}(-0.6, 0.04)$, $\mathcal{N}(-0.2, 0.04)$, $\mathcal{N}(0.2, 0.04)$ and $\mathcal{N}(0.6, 0.04)$, respectively. For the homogeneous population 1, $\boldsymbol{w}$ was drawn from the normal distribution $\mathcal{N}(-0.6, 0.04)$.

For Setting A the values were assigned according to the respective component membership of the areas without error, i.e. for an area i in component k, w_i was set to be the ith value in $\boldsymbol{w}_k$. For Setting B, w_i was drawn randomly from the vector of K candidate values for area i. Figure 5.8 illustrates the resulting histograms of $\boldsymbol{w}$ for one exemplary run in a 2-component scenario. Figure 5.9 shows the respective plots for population 6 with 4 components.

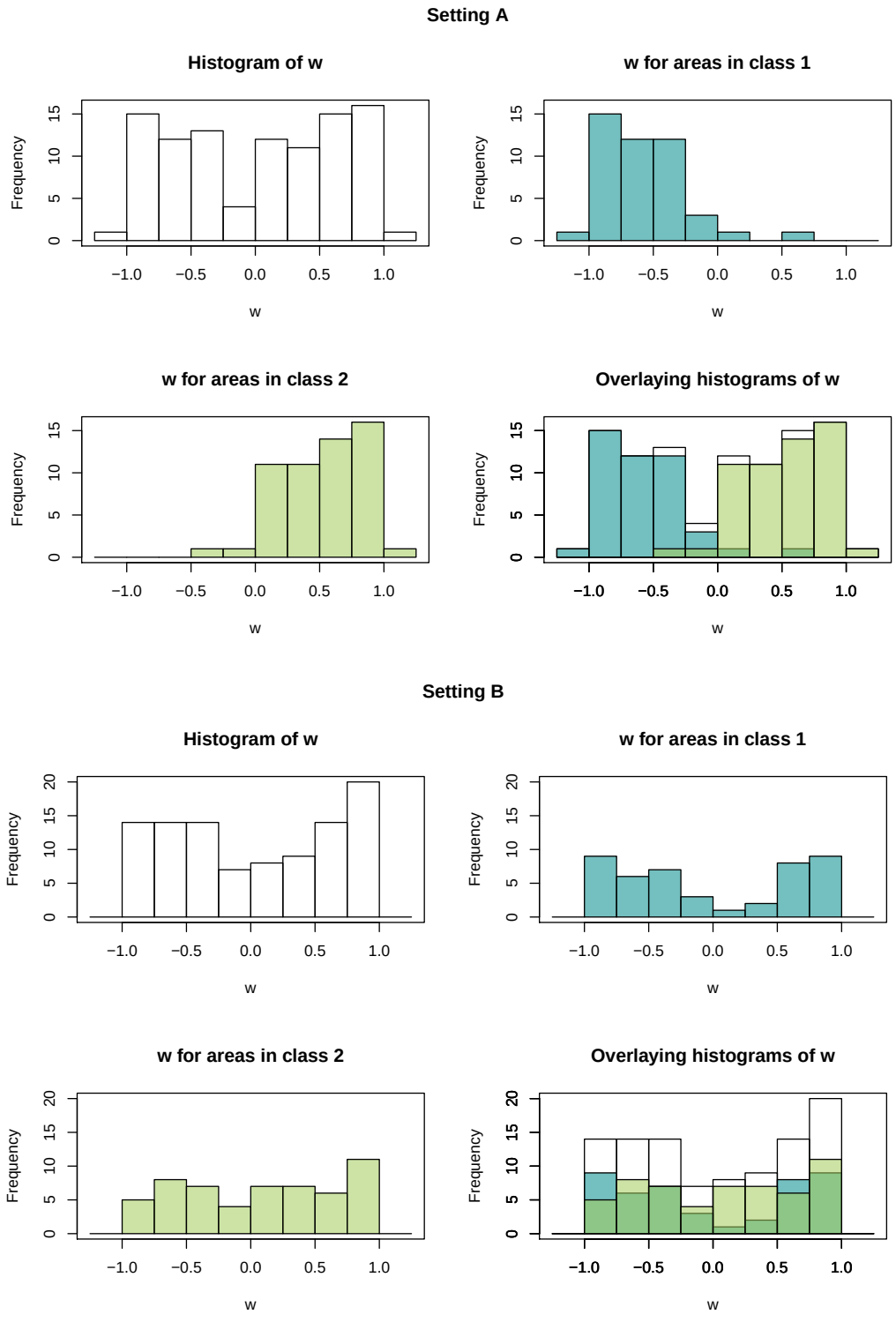

Figure 5.8: Histograms of $\boldsymbol{w}$ for exemplary MC-run ($K = 2$)

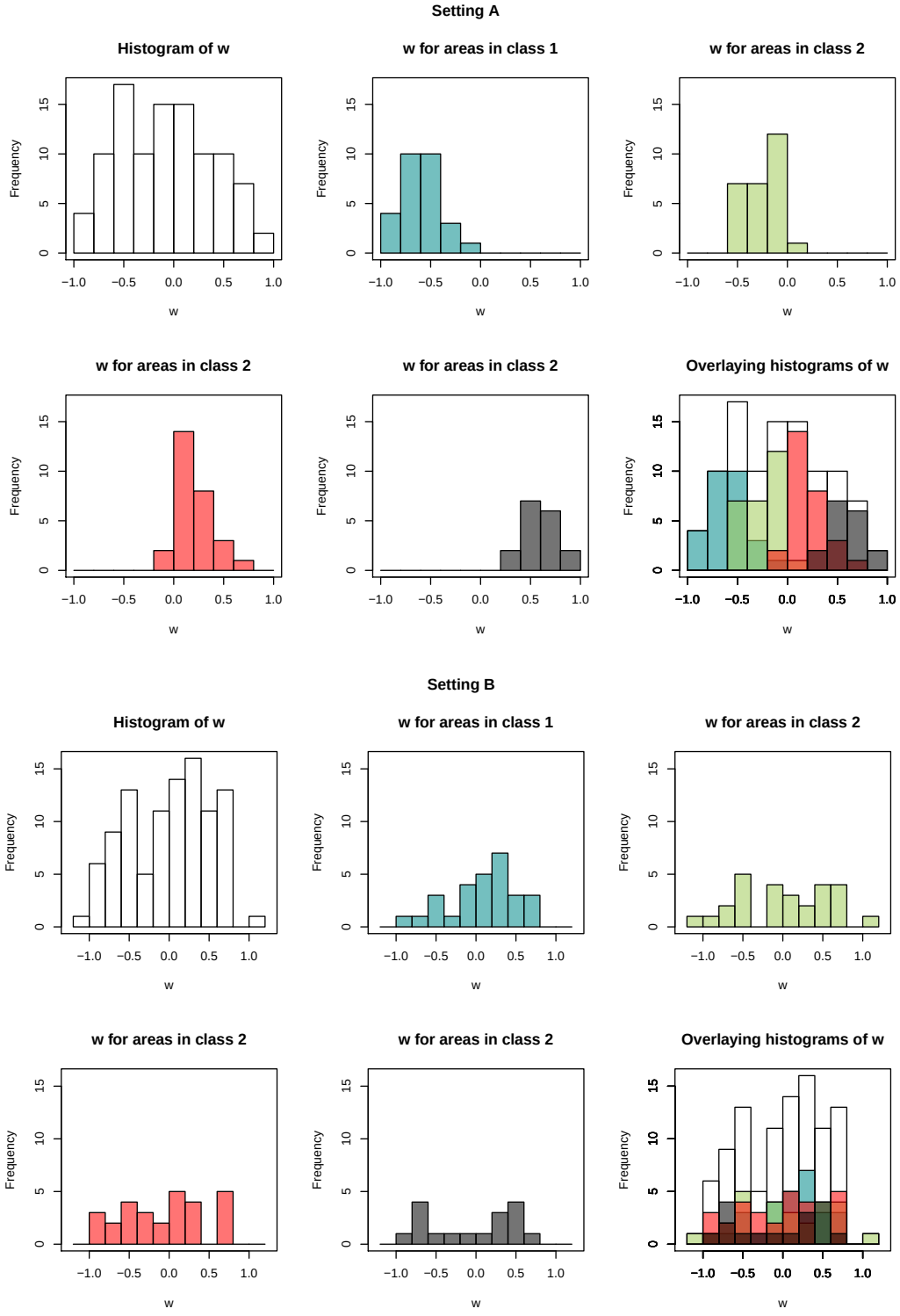

Figure 5.9: Histograms of $\boldsymbol{w}$ for exemplary MC-run ($K = 4$)

As in the area-level study, the observed values in the simulation study are generated from the main model assumed for the component they belong to. On the contrary, the submodel for the mixture weights is conceptualized as an *a-posteriori* tool utilized to understand the data, not as some kind of true data-generating process in the sense that there is a true area-specific mixture weight λ explicitly defined through the submodel. Also note that, as before, $\boldsymbol{X}$ contains no systematic information about component membership.

Table 5.9: Populations in the simulation study

Population 1 homogeneous	*Population 2* clearly separated	*Population 3* partly overlapping	*Population 4* unequal size, clearly separated	*Population 5* unequal size, partly overlapping	*Population 6* highly clustered	*Population 7* different explanatory power
$K = 1$	$K = 2$	$K = 2$	$K = 2$	$K = 2$	$K = 4$	$K = 2$
	$\lambda_k = 0.5 \; \forall \; k$	$\lambda_k = 0.5 \; \forall \; k$	$\lambda_1 = 0.8$, $\lambda_2 = 0.2$	$\lambda_1 = 0.8$, $\lambda_2 = 0.2$	$\lambda_k = 0.25 \; \forall \; k$	$\lambda_k = 0.5 \; \forall \; k$
$\boldsymbol{\beta} = \begin{pmatrix} 6.5 \\ 1.5 \end{pmatrix}$	$\boldsymbol{\beta}_1 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ $\boldsymbol{\beta}_2 = \begin{pmatrix} 9 \\ -2 \end{pmatrix}$	$\boldsymbol{\beta}_1 = \begin{pmatrix} 4 \\ 0.75 \end{pmatrix}$ $\boldsymbol{\beta}_2 = \begin{pmatrix} 5.5 \\ -0.75 \end{pmatrix}$	$\boldsymbol{\beta}_1 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ $\boldsymbol{\beta}_2 = \begin{pmatrix} 9 \\ -2 \end{pmatrix}$	$\boldsymbol{\beta}_1 = \begin{pmatrix} 4 \\ 0.75 \end{pmatrix}$ $\boldsymbol{\beta}_2 = \begin{pmatrix} 5.5 \\ -0.75 \end{pmatrix}$	$\boldsymbol{\beta}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ $\boldsymbol{\beta}_2 = \begin{pmatrix} 4.5 \\ -2.5 \end{pmatrix}$ $\boldsymbol{\beta}_3 = \begin{pmatrix} 6 \\ 3 \end{pmatrix}$ $\boldsymbol{\beta}_4 = \begin{pmatrix} 8.5 \\ -0.5 \end{pmatrix}$	$\boldsymbol{\beta}_1 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ $\boldsymbol{\beta}_2 = \begin{pmatrix} 9 \\ 0 \end{pmatrix}$
$\sigma_u^2 = 1$	$\sigma_{u,k}^2 = 1 \; \forall \; k$					
$\sigma_e^2 = 6$	$\sigma_{e,k}^2 = 6 \; \forall \; k$					
$\boldsymbol{x}_1 = \mathbf{1}$, $\boldsymbol{x}_2 \sim N(4, 0.75)$						

The following estimators were considered:

Table 5.10: Estimators in the simulation study

DIR	Sample mean
BHF	Standard BHF-estimator
clustBHF.mix	Cluster-based BHF estimator. Obtained by (1) clustering of observations based on $\xi_{i,k}$ and (2) subsequent estimation of cluster-specific BHF-models
BHFmixhard	Mixture-based estimator with hard cluster-weights. (Obtained from the model, i most likely belongs to, i.e. $\hat{\mu}_i^{\text{BHFmixhard}} = \sum_{k=1}^{K} \tilde{z}_{ik}\hat{\mu}_{ik}^*$, where z_{ik} is the estimated component label vector (see Section 3.7))
BHFmix	Mixture-based BHF estimator
BHFmixconc	Mixture-based BHF estimator with concomitant variables

Parameter estimation for the mixture-based approaches was performed applying Version 2 of the EM algorithm described in Section 4.3 and 4.6.

Additionally, the performance of different criteria for estimating the number of components K were considered, namely the BIC, its sample-sized adjusted alternative BICadj and the ICL-BIC (compare Section 3.6).

5.3.2 Results

As described in Appendix A.1, the solution of the EM algorithm might depend on the chosen starting values. While the likelihood is never decreased in an iteration of the algorithm, there is nothing to prevent it from converging to a local maximum depending on the initial values. Thus, the algorithm is commonly started repeatedly and the result with the largest likelihood is taken as the final result. With Figure 5.10 convergence issues are analysed for selected populations. The plots show model predicts for area $i = 1, \ldots, m$ for 10 repeated applications of the EM algorithms for different populations and specifications with respect to the assumed number of components. The result with the highest likelihood, i.e. the solution that is returned as result of the algorithm, is indicated by red dots. The

illustration shows, that instability of results due to convergence issues is not a large problem in the present setting. Overall, prediction results are stable. An exemption is population 6, where results get increasingly dependent of the initial values as the assumed number of components grows. This is not only true for the overfitting specification with $K = 5$ but also if the correct number of components is assumed. In a setting like this, it, thus, is recommendable to chose a large enough number of repeated initializations in order to obtain a satisfactory result. It, however, has to be kept in mind, that the sample size for population 6, which is clustered into four components, is the same as for the other six populations consisting of only two or one component. Thus, the average number of data points for each component is much smaller. It can be expected that stability of results increases as the number of data points in each cluster rises.

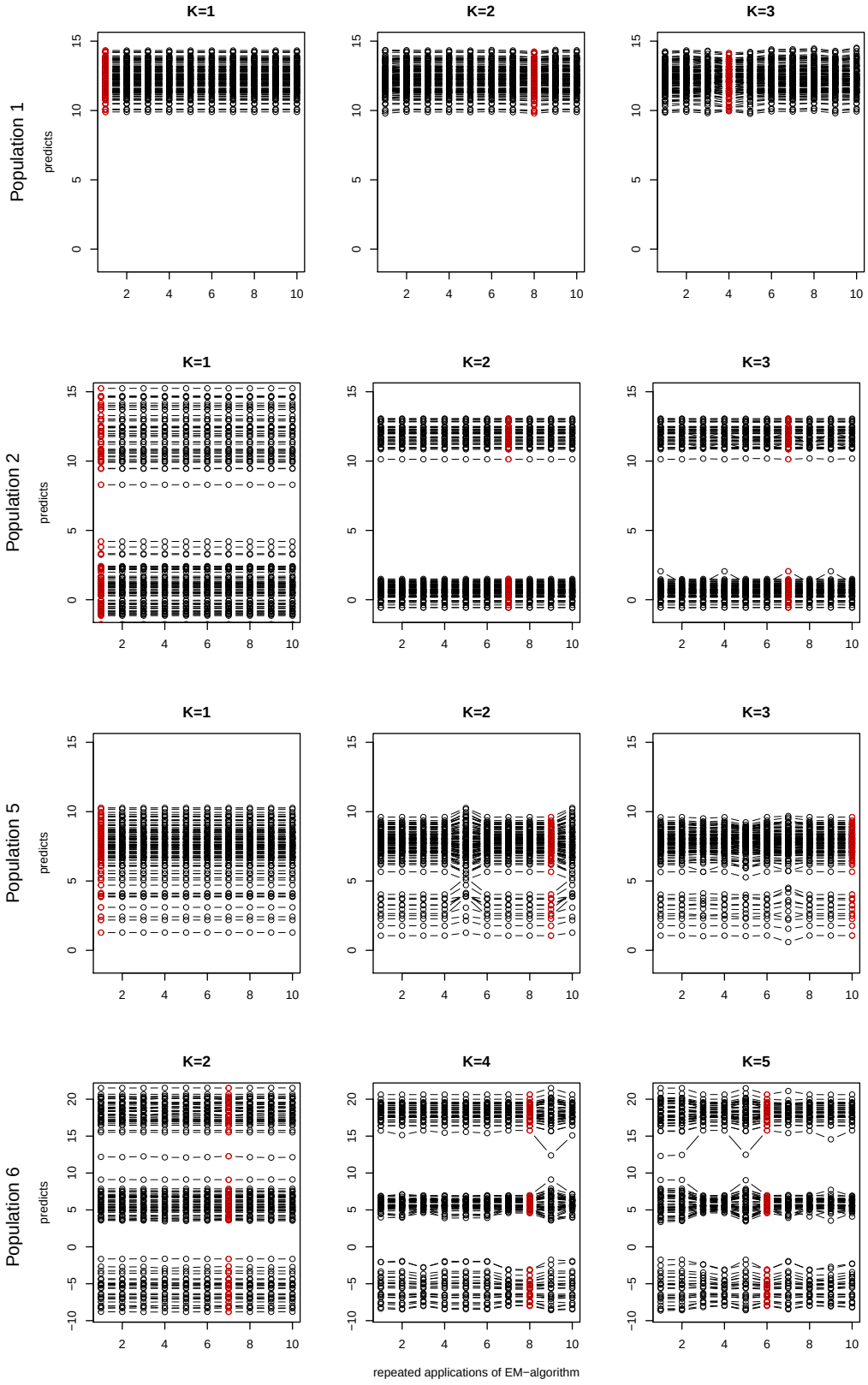

Figure 5.10: Evaluation of convergence

Table 5.11: Simulation results for estimating K

Population	*Criterion*	*Result (Percentage of 1000 MC-runs)*				
		$\widehat{K}=1$	$\widehat{K}=2$	$\widehat{K}=3$	$\widehat{K}=4$	$\widehat{K}=5$
Population 1	ICL-BIC	**100**	0	0	0	0
	BIC	**100**	0	0	0	0
	BICadj	**98.7**	1.2	0.1	0	0
Population 2	ICL-BIC	0	**99.0**	0.9	0 .1	0
	BIC	0	**99.0**	1.0	0	0
	BICadj	0	**96.9**	3.1	0	0
Population 3	ICL-BIC	72.7	**27.3**	0	0	0
	BIC	19.5	**79.9**	0.6	0	0
	BICadj	3.4	**85.1**	10.6	0.9	0
Population 4	ICL-BIC	0	**100**	0	0	0
	BIC	0	**100**	0	0	0
	BICadj	0	**97.3**	2.5	0.2	0
Population 5	ICL-BIC	78.6	**21.4**	0	0	0
	BIC	32.5	**67.3**	0.2	0	0
	BICadj	4.0	**87.9**	7.7	0.4	0
Population 6	ICL-BIC	1.5	0.6	88.1	**7.3**	2.5
	BIC	0.1	2.1	88.0	**7.4**	2.4
	BICadj	0	1.2	85.2	**10.7**	2.9
Population 7	ICL-BIC	95.7	**4.3**	0	0	0
	BIC	24.6	**75.6**	0	0	0
	BICadj	2.1	**94.0**	3.6	0.2	0

In each simulation run the number of components in the population, K, was estimated by calculating the BIC, the sample size adjusted BIC (BICadj) and the ICL-BIC letting K grow from 1 to K (compare Section 3.6). For each population, table 5.11 presents the percentage of MC-runs where the respective estimate $\widehat{K}$ for K was obtained. Results in the column corresponding to the true K are indicated by bold print. It can be seen that for the homogeneous population 1 as well as for the clearly separated populations 2 and 4 all measures estimate the correct number of components with high certainty. As results for population 4 show, this is also true if components are of unequal size. For populations 3 and 5 results for the three measures differ in a way consistent with their properties described in Section 3.6: The ICL-BIC, which penalizes poorly separated components, tends to result in $\widehat{K}=1$, especially if components are of unequal size. The BIC, on the other hand, detects the overlapping components with relatively high certainty and results in $\widehat{K}=2$ in most MC-repetitions. This is particularly true if they

are of equal size, but performance in population 5 is also satisfactory. None of the considered measures is reliable to select the true number of components in the highly clustered population 6. Again it has to be kept in mind that the sample size for this population which is clustered into four components, is the same as for the other six populations consisting of only two or one component. Results might improve considerably if the sample size was risen proportionally. Overall, BICadj is clearly outperformed by the other two measures. Compared to the competing criteria it has a slight tendency to overestimate K and performs worse than the unadjusted BIC in almost all scenarios.

Figure 5.11 and 5.12 illustrate the BIAS and the RMSE for all considered populations and estimators.

As in the area-level simulation in the preceding section, population 1 was introduced to test for the consequences of assuming a clustered structure when the population actually is homogeneous. The same crucial result is obtained: The reduction of RMSE, realized through employing small area estimation techniques, is retained when applying the suggested mixture-based estimators FHmix and FHmixconc in a homogeneous setting. The estimators seem to be robust against this kind of misspecification. This result is obtained both in Setting A and B, i. e. imposing a clustered structure through the submodel in FHmixconc has no hazardous consequences, either. The two-step procedure clustBHF.mix is clearly outperformed. BHFmixhard, too, does yield slightly worse results in terms of RMSE.

Population 2 reveals the performance of the compared approaches in a population that actually is clustered. It is obvious that BHF, while not causing any bias compared to DIR, is not able to reduce the RMSE of prediction. This can easily be explained by considering the estimate for the model variance σ_v^2, which are quite large and result in average shrinking factors $\overline{\gamma}_i^{\mathrm{MC}} = 1/R \sum_{r=1}^{R} \widehat{\gamma}_{i,r}$ ranging from 0.9433 to 0.9574 with mean 0.951. All mixture-based estimators are able to notably reduce the RMSE without causing additional bias: The only exemption is BHFmixconc in Setting A where the inclusion of the submodel, while slightly reducing the RMSE even further, causes a larger bias for few outlying areas. Estimating FHmixconc instead of FHmix, here, does thus not seem to be necessary. The small gain in accuracy does not justify the risk of additional bias. Further, the suggested estimators do not perform better than the alternative two-step approach of clustBHF.mix or the hard-clustering strategy of FHmixhard. Only clustBHF.k is clearly outperformed, even in Scenario A. Thus, judging from population 2 it is recommendable to account for the clustered structure, but there is no evidence that the suggested approaches of BHFmix and BHFmixconc are more suitable than the considered competing estimators. Further, the inclusion of a submodel does

not seem recommendable as the additional further improvement realized through applying BHFmixconc instead of BHFmix is really small.

Population 3, however, sheds light on the performance of the different estimators in the case of less clearly separated components. Here, BHFmix and BHFmixconc outperform the mixture-based alternatives. This is especially true for clustBHF.mix, which is not even able to improve estimation results compared to the benchmark approach of BHF. Further, the inclusion of a submodel, while causing some bias for few areas, now is able to considerably reduce the ARMSE if the covariates are suitable (Setting A). At the same time, the inclusion of unrelated auxiliary information in the model of the mixture weights does not cause any additional bias and does and only slightly diminishes the gain in accuracy compared to FHmix – FMmixconc still performs better than all competing approaches.

Population 4 and 5 repeat the settings of population 2 and 3, with the only difference that clusters are now unbalanced. By and large the same result as before is obtained, i.e. the fact that components are of unequal size has no influence on the performance of the compared estimators. Results for population 6 reveal that the suggested estimators also perform well in a highly clustered scenario.

Overall a strong result is obtained: Different to BHFmixhard and clustBHF.mix, the suggested estimators are robust against misspecification with respect to the actual number of components. At the same time they perform better in a scenario of overlapping components. Moreover, they never perform worse than the benchmark approach of a standard BHF estimator. Summarizing these findings, the suggested approaches flexibly adjust to the data structure at hand, improving estimation results in the case of a clustered population while never deteriorating it. This robustness is a desirable feature of a small area estimator considering that in any real-data-application the true structure is unknown.

Comparing BHFmix and BHFmixconc, it is obvious from population 3 and 5 that estimation results can be further improved through introduction of a submodel that supports subgroup assignment in case of overlapping components. This potential for additional gains in accuracy comes at a rather low price as BHFmixconc generally seems to yield comparable results to BHFmix in Setting A (with the exemption of very few outlying areas in scenario 2 and 4) and also performs reasonably well in case of unfitting covariates. Again, the approach seems to be quite robust against false assumptions regarding the underlying clustering pattern and its determinants.

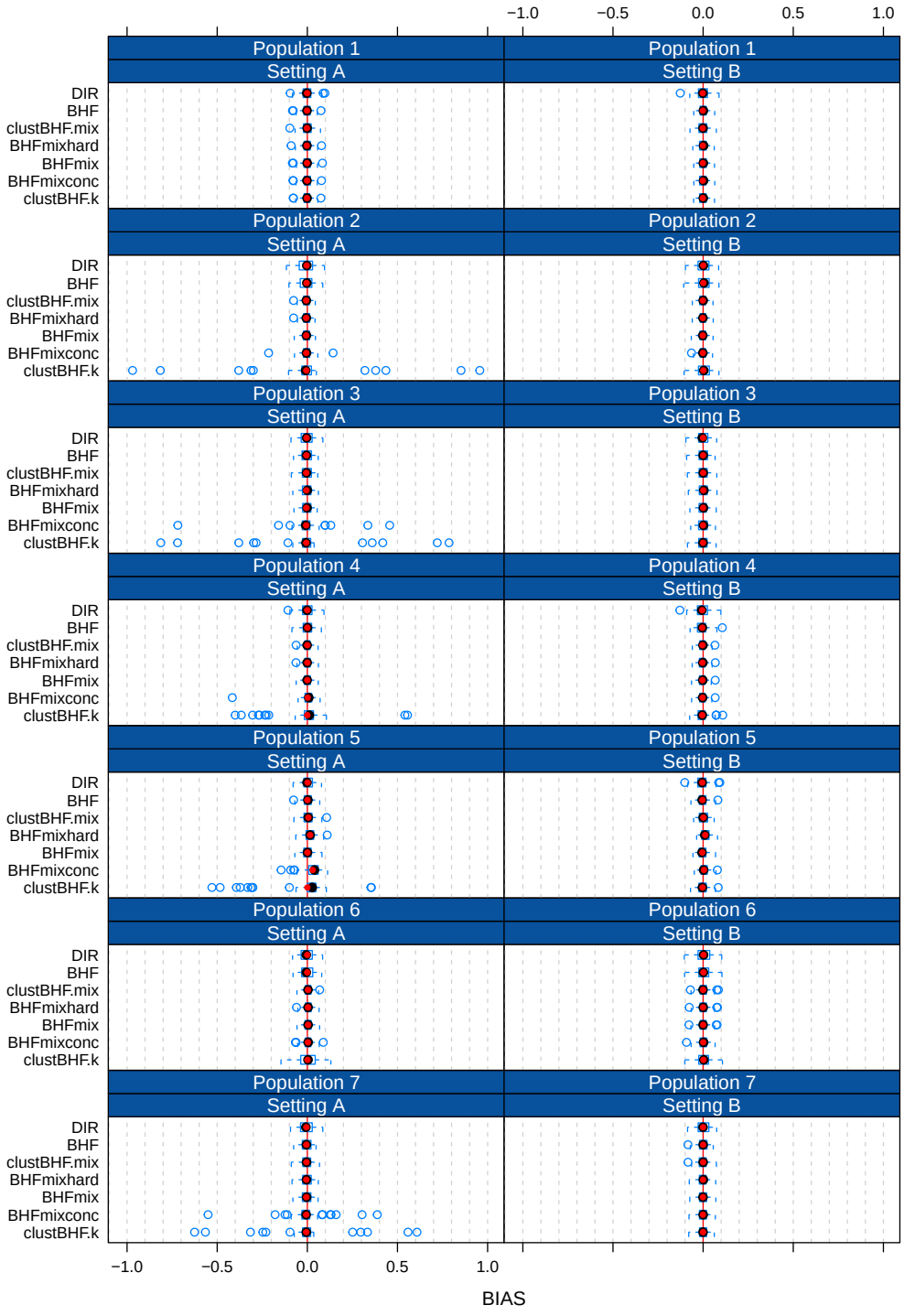

Figure 5.11: BIAS

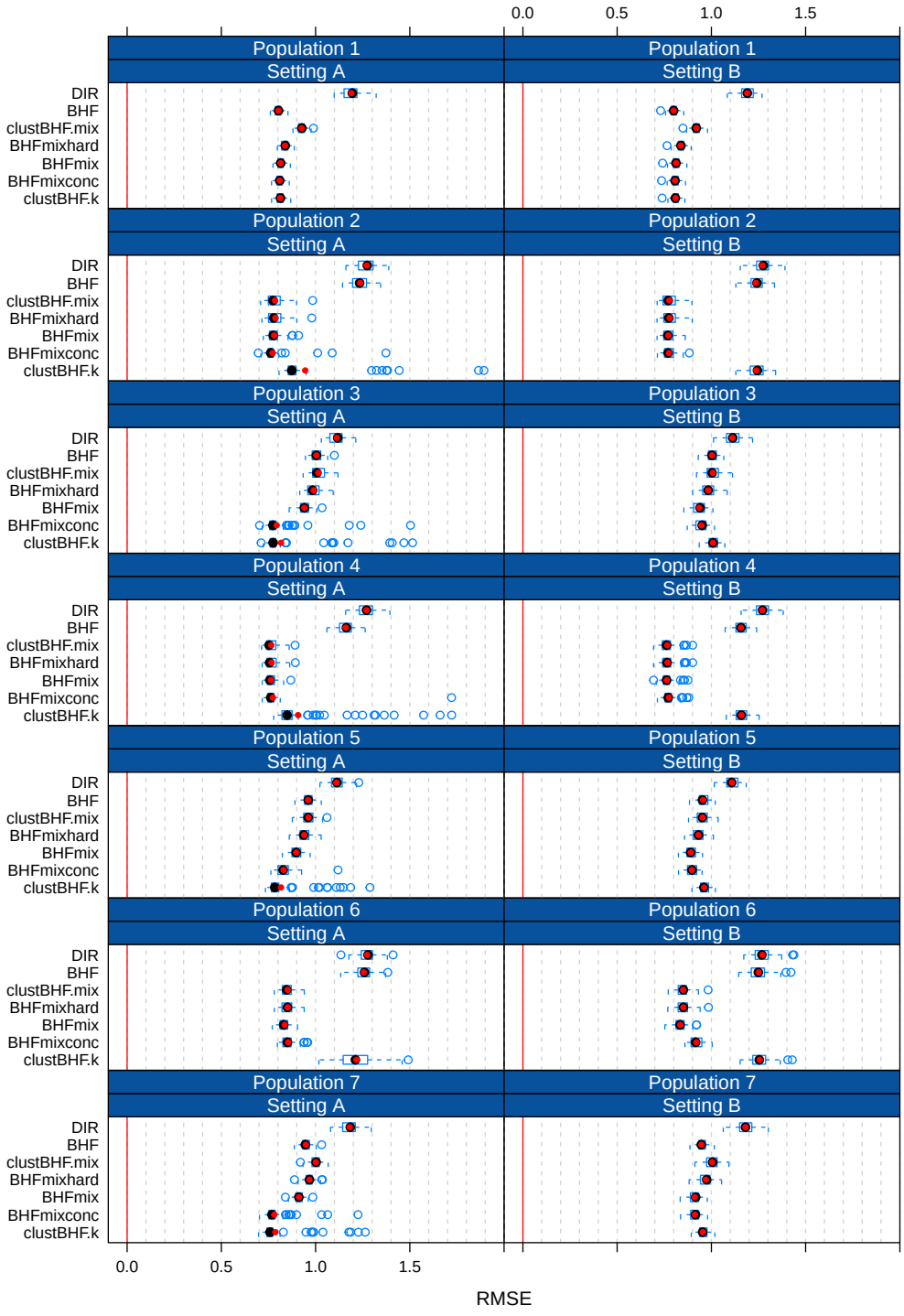

Figure 5.12: RMSE

With Table 5.12 the clustering performance of the proposed estimators is evaluated and compared to the competing approach of clustering via k-means clustering based on $\boldsymbol{w}$. It shows the average percentage of correctly assigned areas over the MC-replications, i.e. mean and standard deviation over 1000 runs. As described in Section 3.7, clustering for BHFmix and BHFmixconc is performed by $k = \text{argmax}_k(\hat{\xi}_{i,k})$. It can be seen that BHFmix and BHFmixconc perform almost equally well in the case of an uncorrelated auxiliary variable for the submodel (Setting B). This substantiates the result already obtained with respect to prediction performance: Using weak or even misleading information in the submodel for the mixture weights does not cause a deterioration of the estimation results. As to be expected, both mixture based approaches clearly outperform clustFH.k, where clustering is solely based on the concomitant variable $\boldsymbol{w}$. Furthermore, the mixture based approaches still outperfrom k-means clustering when strong covariates are available (Setting A). This is not only true for BHFmixconc, that uses the supplementary information, but also for BHFmix. The only exemption is population 7, where the assumed main model is misspecified for one of two components. The comparison between BHFmix and BHFmixconc reveals that the submodel for the mixture weights supports clustering in cases where the components are not clearly separated. This can be seen in the results for population 3,5 and 7, where the average percentage of correctly assigned areas rises when including concomitant variables.

Table 5.12: Percentage of areas correctly assigned to clusters: Mean and standard deviation over MC-runs

		Setting A		*Setting B*	
		mean	std.dev	mean	std.dev
Population 2	BHFmix	99.72	1.65	99.82	1.36
	BHFmixconc	99.84	0.50	99.71	1.73
	BHFclust.k	96.34	1.53	50.94	4.84
Population 3	BHFmix	85.29	20.03	86.26	19.48
	BHFmixconc	89.49	27.94	86.54	14.51
	clust.k	84.99	30.43	50.38	4.91
Population 4	BHFmix	99.95	0.22	99.73	4.48
	BHFmixconc	99.67	0.65	99.63	0.93
	clust.k	90.44	13.77	51.03	4.84
Population 5	BHFmix	80.18	28.60	80.81	28.00
	BHFmixconc	91.42	13.74	83.37	12.25
	clust.k	76.01	33.88	50.56	4.92
Population 7	BHFmix	85.04	6.70	85.23	6.43
	BHFmixconc	97.62	1.80	85.45	5.84
	clust.k	96.34	1.53	51.15	4.79

Chapter 6

Application: Estimating Rental Prices for German Districts

The suggested method was motivated by the intention to estimate rental prices for Germany at the district level (NUTS-3). Direct estimates for this purpose were obtained from the German *Mikrozensus 2010*, a 1%-household survey conducted by the Federal Statistical Institute (Statistisches Bundesamt (2011), Statistisches Bundesamt (2012)). The *Mikrozensus* is the only nation-wide survey providing estimates of regional rental prices. It moreover delivers these information on a regular basis so that it is a valuable source of information. Albeit being the largest regular household survey in Germany, the *Mikrozensus* is however not designed to be evaluated at a regional level (Statistisches Bundesamt, 2011). Accordingly, results on average rental prices are only published for the 16 German Länder and, in some cases, for 38 regions (NUTS-2). For the application at hand, a special evaluation was provided on a far stronger regional disaggregation level: It contained average rents per square meter at the district level for 13 of the 16 German Länder, i.e. for 246 of 412 districts. Additional to the average rental prices, district-specific sample sizes n_i as well as the estimated design variances of the direct estimates were provided. As frequently done in practical applications, it was assumed that $\hat{\sigma}^2_{e,i} = \sigma^2_{e,i}$, i.e. the estimates were used as the presumably known variances of the direct estimates. This implies ignoring the variability of the estimates (For a discussion of the implications of this assumption, see Bell (2008)). Auxiliary information was obtained from a broad range of regional indicators on district level provided by official statistics in Germany and openly available at `http://www.inkar.de` (see Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR), 2017).

Direct estimates were based on a total of 112.142 observations, with area-specific

sample sizes ranging from 51 to as much as 12.008 units (average sample size $\bar{n}_i = 455.9$). Accordingly, estimated coefficients of variation (CVs) range from 0.007 to 0.25, with mean 0.06. Only 18 areas have a CV larger than 0.10. Figure 6.1 shows area-specific sample sizes and the estimated standard deviations against estimated rental prices. The large variation in sample sizes and consequently in CV clearly reveals, that the design of the survey was not optimized for an evaluation on district level. Of course, CVs for most areas are nevertheless atypically low for an application of small area techniques. As stated above, the Federal Statistical Office does, however, not publish design-based results on district level, claiming that they do not fulfil precision requirements. Small area estimation, thus, seems to be a natural solution.

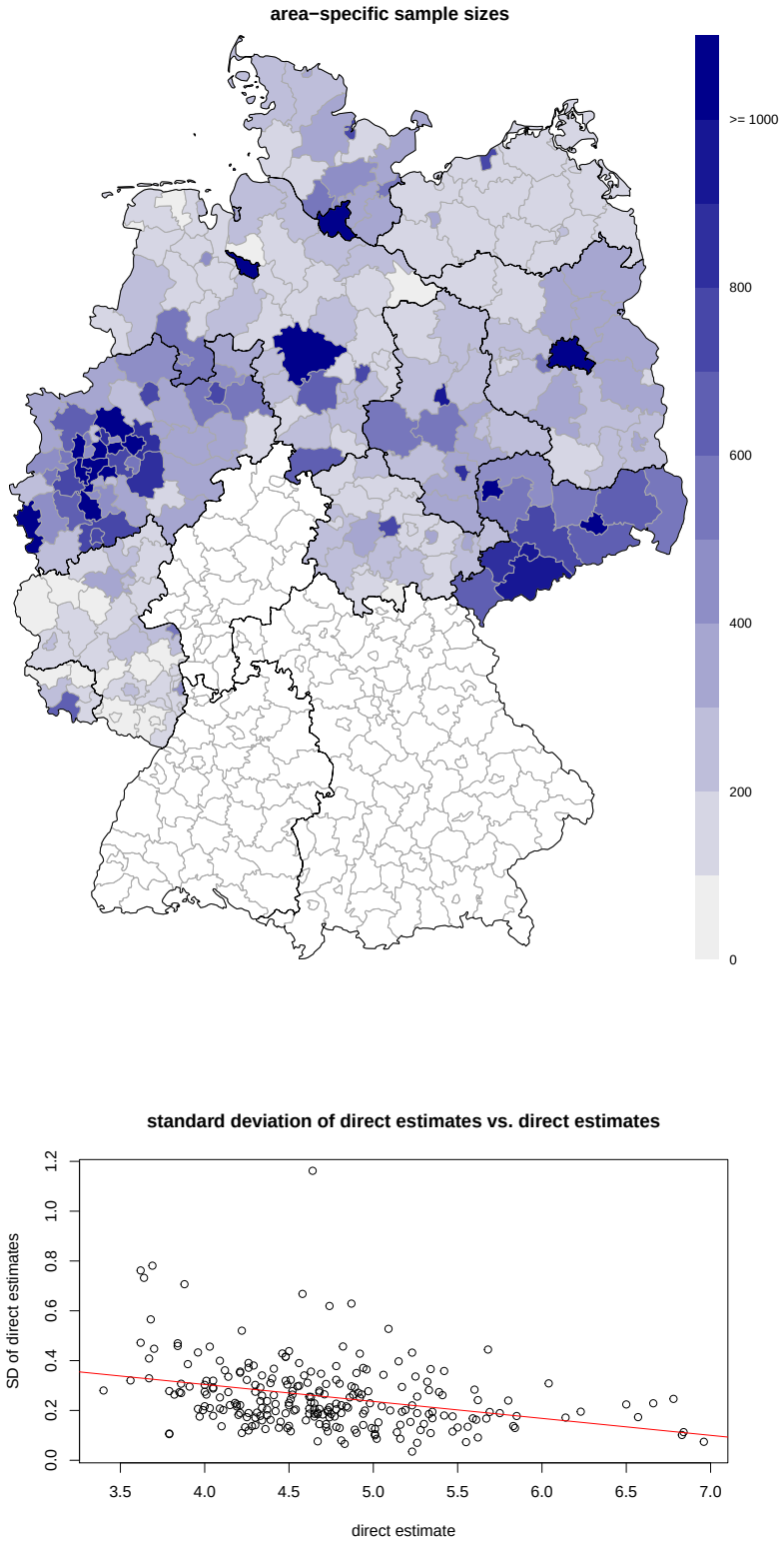

Figure 6.1: Sample sizes and standard deviation of direct estimates

Starting point for the application was the estimation of a standard FH-model (see Articus, 2014). Variable selection was performed by pursuing a literature-based analysis of important driving factors of rental prices as well as by applying simple stepwise selection procedures. As model selection criterion the conditional AIC as suggested by Vaida and Blanchard (2005) was employed. Based on the results, a model including six indicators, namely population growth rate (PGRO), prevalence of rented housing (RENT), vacancy rate (VACQ), employment rate (EMPL), net migration rate (MIGR), price of building land (LAND)) was chosen. See Appendix A.3 for details on all auxiliary information considered in this study. As stated above, the assumption of a common fixed part of the model for all districts drawn in the standard model, however, seemed inappropriate in the application of estimating regional rental prices. When presenting the model to practitioners it was decisively rejected. Instead, they argued convincingly that one should differentiate between rural and urban areas. This criticism motivated the proposal of the estimators suggested in this thesis as well as the search for adequate procedures to test for the existence of subgroups in the population. Building upon the above mentioned study, now the suggested mixture-based estimators are applied for the estimation of regional rental prices.

To assess the number of clusters, both the ICL-BIC and the BIC were considered. Note that this includes the essential decision between $K = 1$ or $K > 1$, i.e. the question whether it is appropriate to assume the existence of latent subgroups and to accept the larger complexity of employing a mixture model in the first place. While the ICL-BIC resulted in $\widehat{K} = 1$, the BIC suggested two clusters. Regarding the results from the simulation study (and corresponding to the features of the two criteria discussed in 3.6) this indicates the existence of poorly separated but nevertheless existent clusters. Therefore, the suggested approach was adopted and a finite mixture of Fay-Herriot models (FHmix) with $K = 2$ and the same set of covariates as in the standard model for both components was estimated. Further, to support subgroup assignment and gain insights into the clustering structure, FHmix was extended to include covariates for the mixture weights (FHmixconc). Because of the motivating notion that determinants of rental prices vary between rural and urban areas, settlement density (DENS) was used as covariate for the submodel. As reference approaches, the standard Fay-Herriot model (FH) as well as the spatial extension of the Fay-Herriot model (FHS) (see Molina, Salvati and Pratesi, 2009; Pratesi and Salvati, 2008) were considered. Additionally, a mixture model with DENS as further covariate in the main model was estimated as a competing approach to FHmixconc.

Estimated coefficients are given in Table A.2 in Appendix A.3.2. It can be taken from these results, that the estimated model variance σ_v^2 can clearly be reduced with the more complex approaches: While it is 0.081 for the FH, it is only 0.039 for

FHS. All mixture-based approaches result in two components with quite different design variances $\hat{\sigma}^2_{v,k}$. Component 1 has an estimated design variance of 0.033 for FHmix, 0.025 for FHmixconc and 0.036 for FHmixDENS. These values are quite close to the result obtained with FHS. There is, however, a second component, i.e. component 2, with a model variance that is considerably smaller: $\hat{\sigma}_{k,1}$ is 0.0055 for FHmix, 0.0069 for FHmixconc and even 1.74×10^{-7} for FHmixDENS. For all estimators, the estimated prior probability for this second component is slightly larger than for component 1. It is $\hat{\lambda}_2 = 0.59$ for FHmix and $\hat{\lambda}_2 = 0.62$ for FHmixDENS. In the case of FHmixconc there are area-specific prior probabilities $\hat{\lambda}_{i,2}$ ranging from 0.08 to 0.88 with mean $\overline{\lambda}_2 = 0.69$.

Smaller estimated model variances, of course, imply that model-based estimators which are a convex combination of a direct and a synthetic estimator (see e.g. Section 2.4) rely more heavily on the synthetic estimator. Figure 6.2 shows the resulting distribution of shrinkage factors $\widehat{\gamma}_i$ for the competing estimation approaches. For the mixture-based approaches in the lower two plots of the panel the distribution of the component-specific shrinkage factors $\widehat{\gamma}_{i,k}$ are depicted. Because area-specific estimates are convex combinations of predicts from all component models, additionally, boxplots for the weighted average $\sum_{k=1}^{K} \widehat{\xi}_{i,k}\widehat{\gamma}_{i,k}$ are depicted. This can be interpreted as the relevant weight of synthetic estimation in the resulting mixture-based estimator (compare Section 2.4). It is obvious from this plot that the overall reliance on the model is larger for the mixture-based approaches. Given the larger flexibility of the modelling approach, this is an expected result.

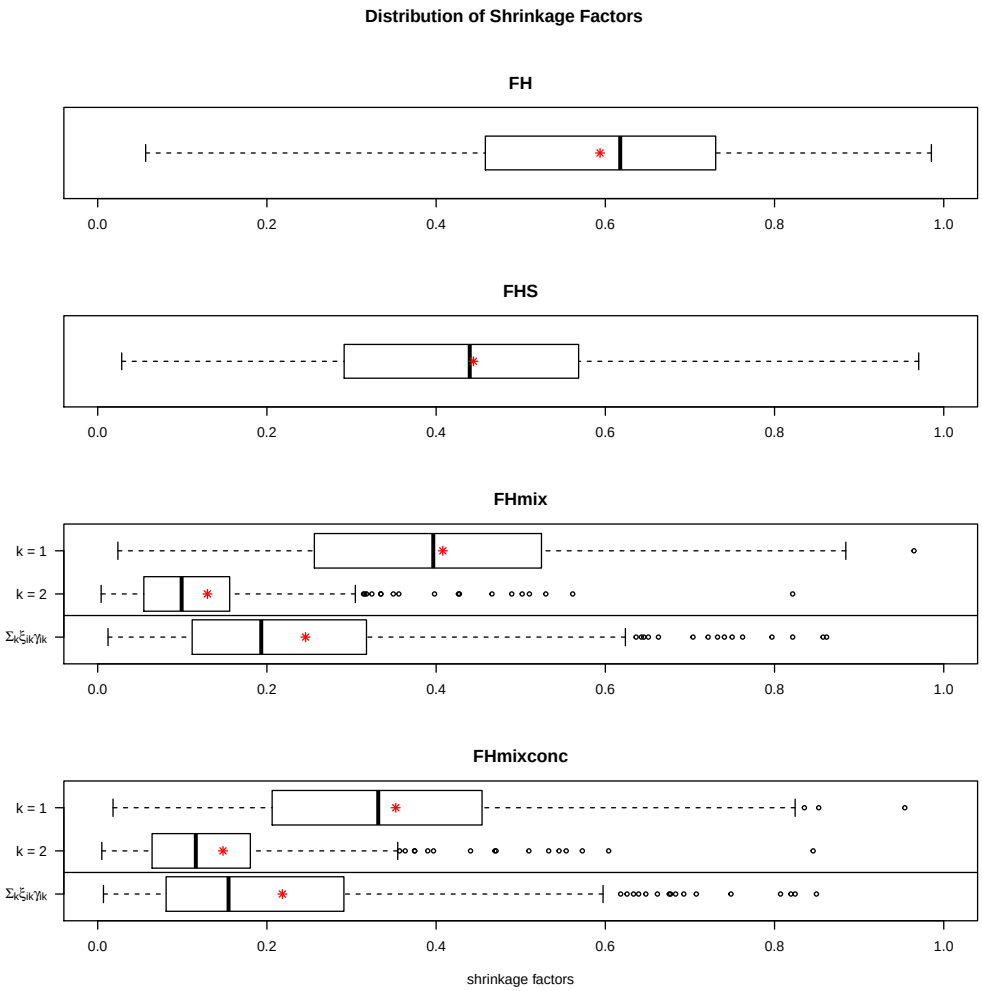

Figure 6.2: Shrinkage factors

In Figure 6.3 small area estimates from all competing models were plotted against the available direct estimates. Thus, the unbiased but imprecise direct estimates obtained from the sample are deployed to judge the bias of model-based estimates. This simple plot has been suggested as a tool for bias diagnostic by Brown, Chambers, Heady and Heasman (2001). To further analyse the cluster-based estimators, points are coloured according to their assignment to one of the two component- or cluster-models. For the mixture-based estimators, blue and red corresponds to a strong conditional probability of belonging to component 1 or 2, i.e. to $\hat{\xi}_{i,1} \approx 1$ and $\hat{\xi}_{i,1} \approx 0$, respectively. Shades on the range between these to colors accordingly indicate values for $\hat{\xi}_{i,1}$ on the scale between 1 and 0. For the cluster-based estimator FHclust, colors indicate the corresponding result of

hard clustering. To complement the findings by supporting the comparison of the competing approaches, a scatterplot matrix of estimation results for all model-based estimators is provided with Figure 6.4. The plots indicate that all models tend to overestimate rental prices on the lower tail of the distribution. This is true for the reference approach of FH and even more pronounced for the mixture based estimators. Note, however, that sample sizes for the low-priced, usually small and rural regions, are small. As Figure 6.1 shows, there are particularly large design variances of direct estimates in these regions and direct estimates are of limited reliability. FH and FHS also seem to systematically underestimate prices in the highest-priced regions. Here, the performance of the mixture-based estimators is clearly better.

A natural question arising with the application of the the mixture-based estimator is whether the estimated conditional probabilities of subgroup membership hint at the existence of two meaningful clusters of areas. Further, it is of interest whether these results change if a submodel for the mixture weights is assumed. To approach these questions, the estimates for $\hat{\xi}_{i,1}$ for FHmix and FHmixconc were plotted in a map (see Figure 6.5). As $\hat{\xi}_{i,1} + \hat{\xi}_{i,2} = 1$, the illustration of conditional probabilities for model 2 is redundant. The spatial representation of $\hat{\xi}_{i,1}$ reveals some kind of an agglomeration effect: Both the districts around Hamburg and the Rhineland-region with cities as Köln, Bonn, Düsseldorf and the cities of the Ruhr region are strongly assigned to model 1. The same is true for cities such as Bremen, Hannover, Münster, Osnabrück, Kiel, Mainz, Rostock, Cottbus and Kaiserslautern. Exemptions are Dresden, Leipzig, Bielefeld and – most strinkingly – Berlin, which is strongly assigned to component 2. At the same time, rural areas mostly have a low estimated conditional probability of belonging to this model. A noticeable exemption is the Vogtlandkreis, a rural district in Saxony. Despite these exemptions, the model indeed seems to roughly differentiate between rural and urban regions. FHmixconc further supports this distinction mainly in the sense of resulting in conditional probabilities that are closer to 0 or 1 and thus less ambiguous. Only very few areas change their subgroup assignment, among them Leipzig and Kaiserslautern, which now are strongly assigned to model 1, too.

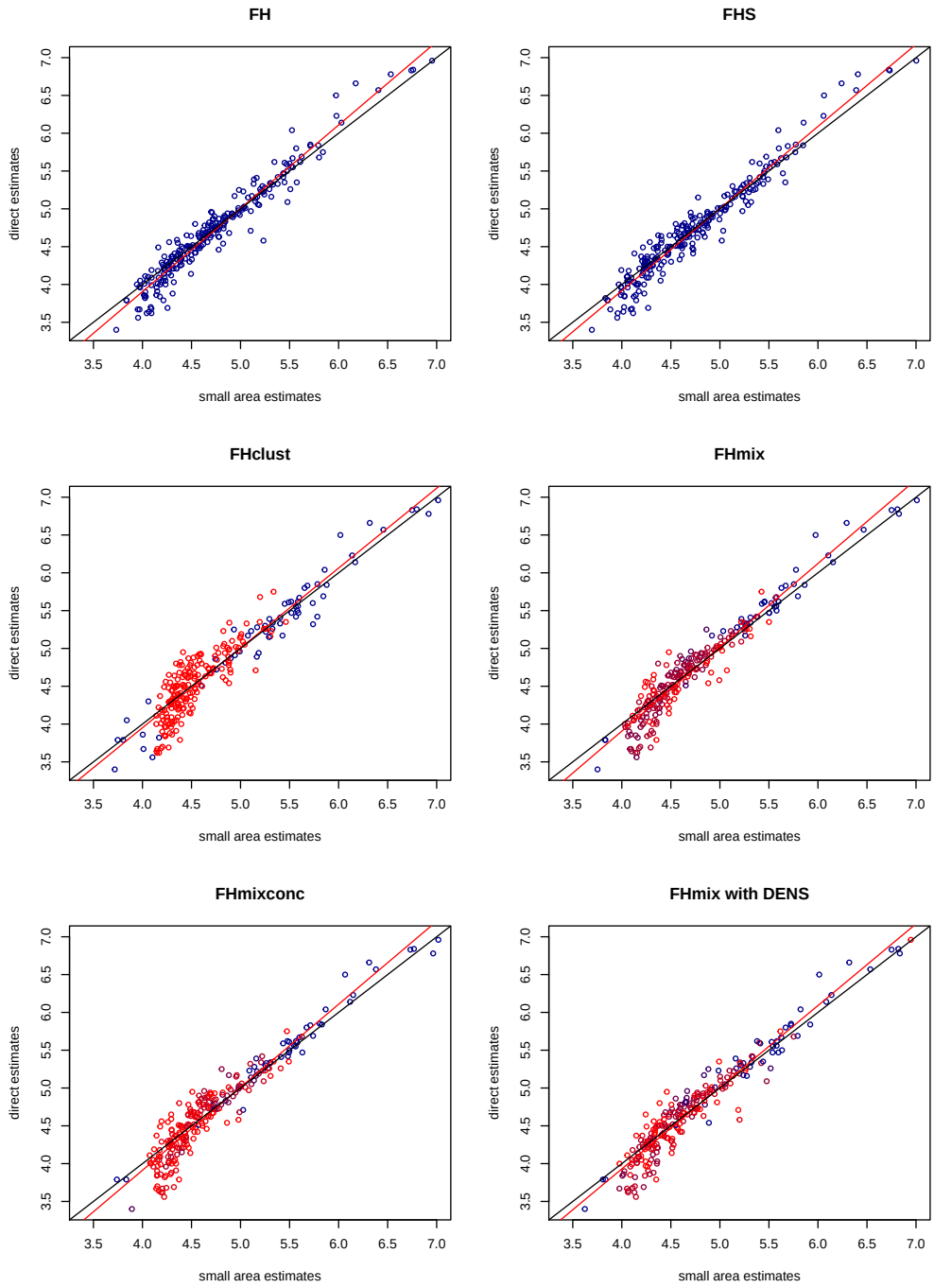

Figure 6.3: Comparison of direct estimates and model-based estimates

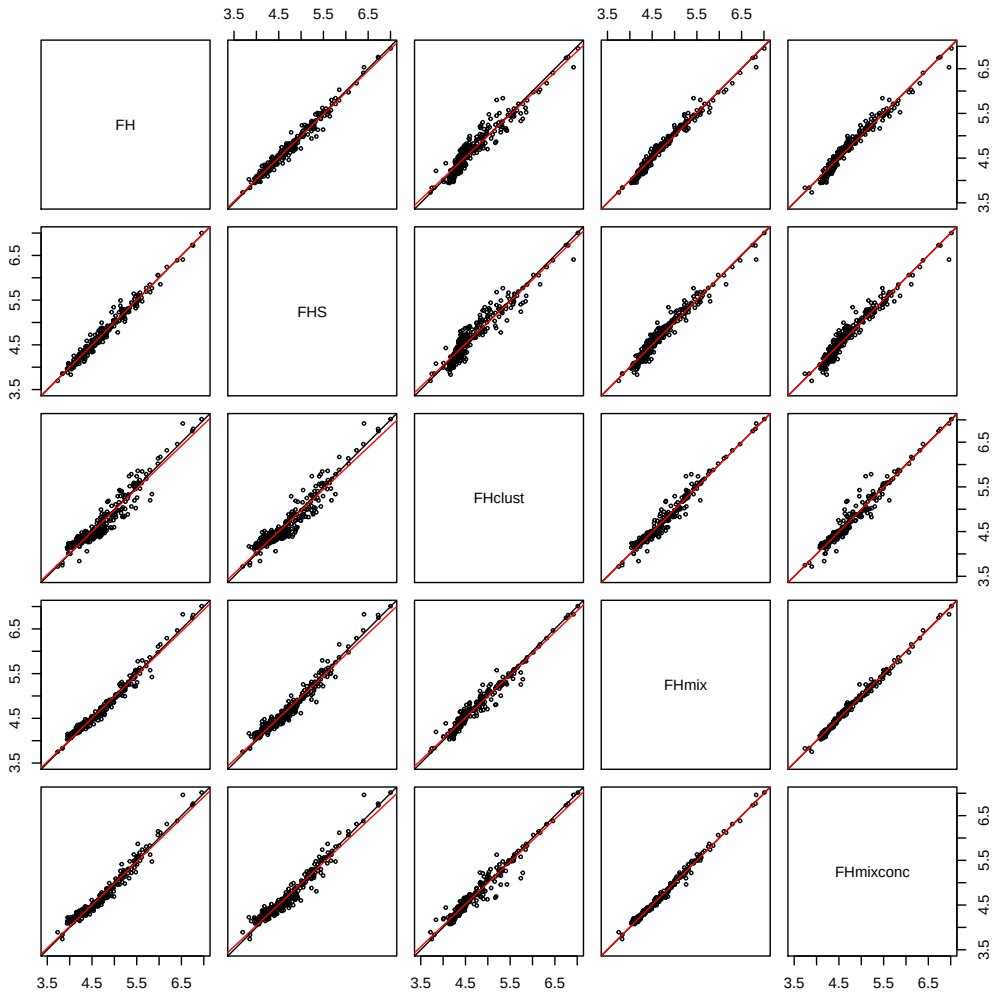

Figure 6.4: Results from competing estimators

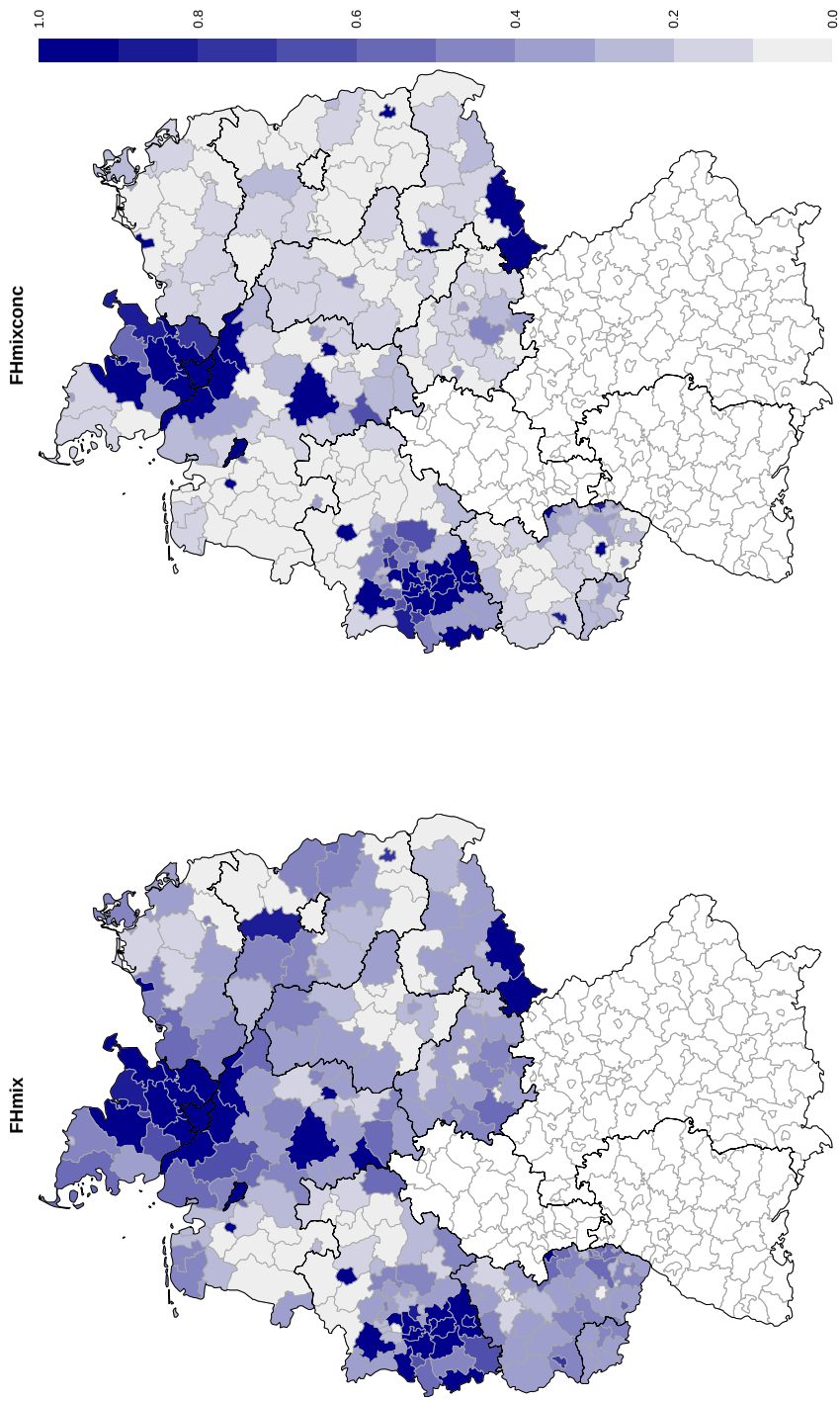

Figure 6.5: Spatial representation of $\hat{\xi}_{i,1}$

The aim of applying methods of SAE is to realize a gain in accuracy in the context of small subsamples and, hence, to overcome large standard errors of traditional direct estimates. With Figure 6.6 the performance of the suggested approach is evaluated in this regard. It depicts boxplots for the distribution of the estimated RRMSE of the proposed mixture based estimates and the standard and spatial FH estimates as well as of the Coefficient of Variation (CV) of the direct estimates. The asterisks mark the respective average RRMSE and CV over the areas. As in the simulation study, the MSE for the mixture based estimators was obtained as the approximation suggested in Section 4.7. The plot shows that a significant gain in accuracy can be realized when applying SAE methods instead of direct estimation. The more complex spatial extension of the FH model yields better results than the standard model. Comparing the proposed estimator and the model-based reference approaches, a considerable further improvement can be made for almost all areas when applying mixture based estimators. The estimated RRMSE for FHmixconc is even smaller than for FHmix. This is in accordance with the finding from the analysis of shrinkage factors, which showed that the overall reliance on the synthetic part of the estimator is particularly strong for this estimator. It is however important to bear in mind that the suggested MSE approximation for the mixture based estimators seems to underestimate the true MSE. Further research is necessary to develop a more reliable uncertainty measure.

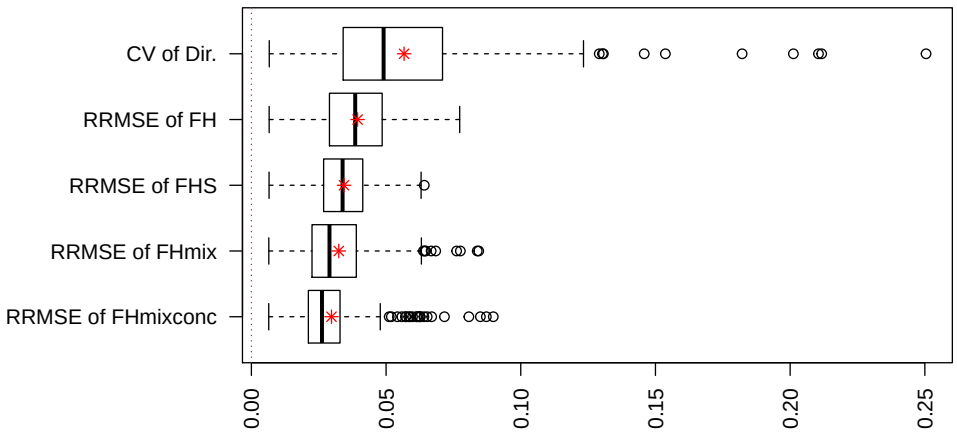

Figure 6.6: Estimated RRMSE for competing estimators

Finally, the results for rental prices on district level for 2010 obtained from both FHmix and FHmixconc are illustrated in Figure 6.7. Results for FHmix and FHmixconc are similar with the only exemption, that FHmix yields a slightly higher result for some of the lower-priced regions in Lower Saxony and Mecklenburg-

Vorpommern. Figure A.7 in Appendix A.3.3 contains a spatial representation of quoted rents on district-level for the year 2011, published by the Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) (2012) based on the study already mentioned in the introduction. While quoted rents have some methodological shortcomings and are generally higher than those actually paid by residents, they can be expected to be highly correlated with the values obtained here. They can, thus, be employed for external validation of the estimation results. The comparison of the maps shows that regional patterns indeed are quite similar, the only striking exemption being the region surrounding Berlin.

Estimated prices range from approximately 3.70 to slightly more than 7.00 Euro per square meter. As expected, the map clearly shows the particularly high prices in large cities such as Hamburg, Köln, Düsseldorf, and Mainz. It also reveals that the price levels in these cities also affect surrounding regions. Rural districts in Eastern Germany, especially Saxony-Anhalt and Saxony, and some areas in Rhineland-Palatinate and Lower Saxony are identified as especially low-priced. The example of Berlin and the very large districts in Brandenburg show that the level of analysis still is to large to adequately represent rental price levels. A single price level for Berlin gives the misleading impression of moderate prices in the city. All the same, the averaging of results for districts in Brandenburg masks the influence of the capital's rental market on the surrounding regions. This is revealed by comparison with the BBSR-study that in these districts differentiates between an infra-structurally integrated area and a less integrated region. It is obvious from this example and also from registering the large differences of the layout of districts between federal states that the considered districts are historically grown entities of administration with regional particularities and not areas designed for statistical analysis.

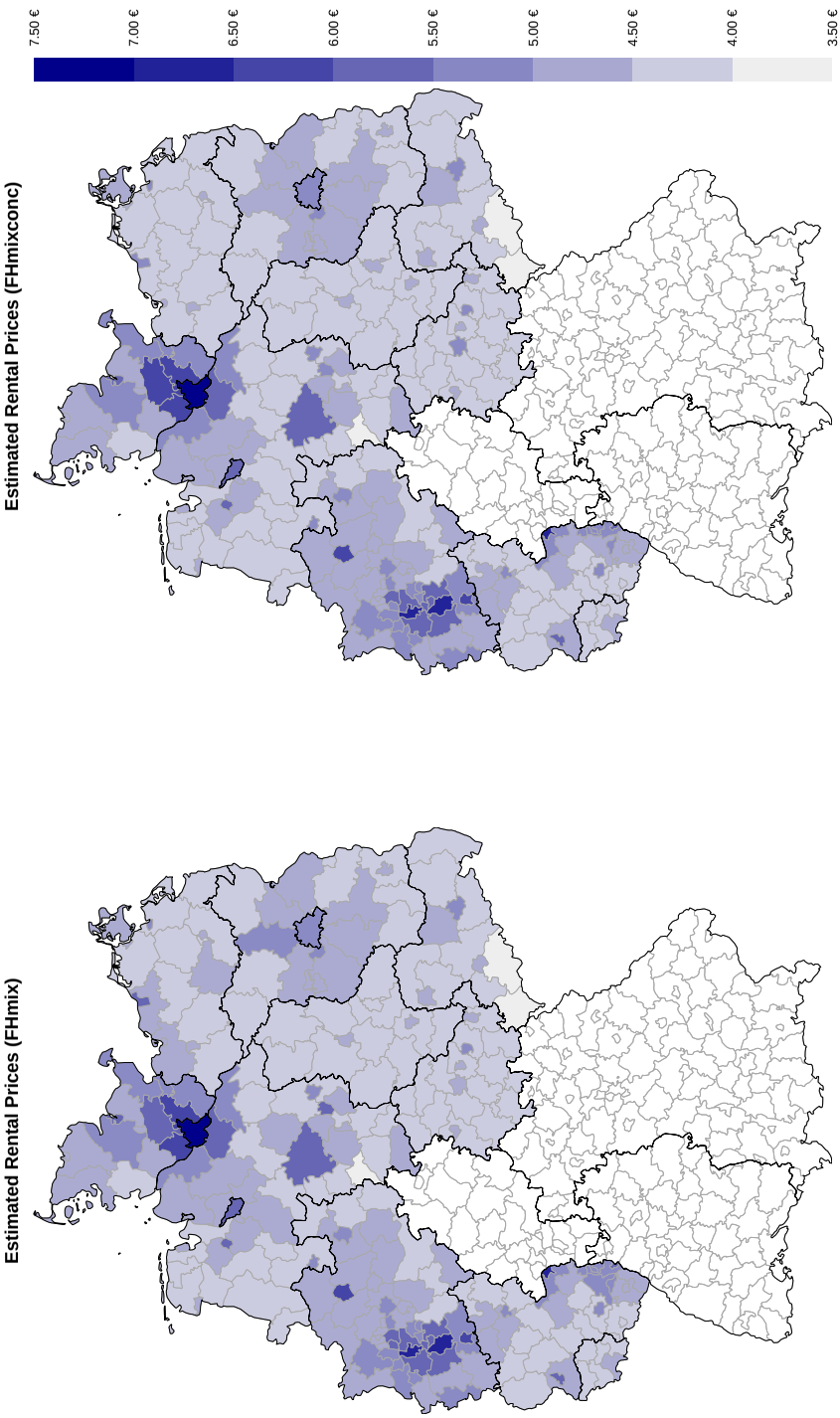

Figure 6.7: Estimated rental prices (FHmix and FHmixconc)

Chapter 7

Conclusion

In this thesis, mixture-based small area models were suggested to account for the existence of unobserved subgroups within the population. Based on an account of relevant theory from the fields of SAE and FMM, both a finite mixture of unit-level and area-level models were proposed as special cases of a mixture of linear mixed regression models. In addition, the models were extended to include (concomitant variable) submodels for the mixture weights. At the same time, the framework was transferred into the specific language and notation of SAE, and the particularities and distinct focus of the discipline were accounted for. Estimation of model parameters was discussed alongside relevant criteria for estimating the number of components. Finally, a mixture-based estimator for SAE was derived in the form of a plug-in estimator based on the BP for a suitably formulated true area mean. This estimator predicts the statistic of interest as a weighted average of the predictions from the component-models. To assess the prediction error, an approximation of the conditional MSE was suggested. The proposed estimators were evaluated in two model-based simulation studies and then applied to the problem of estimating regional rental prices in German districts.

Simulation studies were conducted to analyse the performance of the suggested estimators under different scenarios. They showed that the mixture-based estimators are indeed able to improve estimation accuracy in cases of clustered populations and, in that, overall outperform competing cluster-based approaches. Furthermore, they never demonstrate performance that is inferior to that of standard models, even if the assumption of a clustered population is false, i.e. the true number of components is one. This robustness against misspecification is a strong feature of an estimator that is intended to be applied in a real-data-application, in which the true structure of the data to be investigated remains unknown. The simulations also demonstrated that the estimators' performance can be further

improved by including suitable concomitant variables for modelling the mixture weights. This is particularly true if the main model components are not well separated, so that there is a potential or need to support the clustering. Again, the approach is strikingly robust against misspecification in the sense that imposing a false clustering structure through the submodel does not seem to negatively influence estimation results. Finally, the performance of the suggested uncertainty measure was evaluated. Simulations showed, that it tends to underestimate the true MSE and thus indicated the need of further research.

The suggested area-level estimators were employed to estimate regional rental prices in German districts. While the study revealed that the layout of these districts is not optimal for an analysis of rental price levels, plausible results could nevertheless be obtained. Overall, model variance was considerably reduced by employing the flexible mixture models. This, corresponding to the trade-off generally faced in model-based SAE, comes at the price of some additional bias. Analysis of the clustering structure indicated some kind of agglomeration effect, confirming the existence of rural and urban particularities that initially motivated the proposal of a mixture model.

In summary, the results are promising. Most importantly, the suggested estimator is indeed able to improve estimation performance in cases involving unobserved subgroups. The model selection criteria discussed for the choice of K function reasonably well, thus allowing researchers to assess this crucial question of model specification in practical applications. Moreover, as a by-product of the estimation process, the suggested approach yields area-specific probabilities of subgroup assignment that can be employed to partition areas into clusters. In addition, probabilistic subgroup assignments provide further insights into underlying data structures and help to understand the data situation at hand. Given that the mixture-based approach is also intuitively appealing, it may prove to be an attractive estimation strategy in any application in which areas are suspected to be divided into a number of latent subgroups. And while, in this thesis, the use of mixtures was motivated by the existence of actually existent subgroups, the suggested approach can, of course, also be used in settings in which components do not correspond to clusters that are existent in some physical sense. As McLachlan and Peel (2000) emphasize, mixture models can be interpreted more broadly as a framework to flexibly account for heterogeneity in a population or to semi-parametrically model unknown or unsmooth distributional shapes. This opens up a wide range of possible uses in SAE.

It thus seems worthwhile to pursue the approach further. Future research might include a more thorough analysis of the estimators as well as complementing them with tools intended to facilitate their use in real-data applications. More particu-

larly, the following questions are regarded as important tasks for future work:

To complement the findings of the simulation study and to further analyse the features, strengths and weaknesses of the suggested estimators, a design-based simulation study, set in a close-to-reality setting, should be conducted. This might reveal any unexpected peculiarities of the estimators and, thus, help to better understand their functioning and judge their suitability in real-data applications. Building on the results of the simulations, a further central topic of future research is assessing the properties of the mixture-based estimators theoretically. In addition to the standard catalogue of relevant properties, it might also be interesting to theoretically analyse the robustness of the estimators against misspecification with regard to the assumed number of components and the submodel for the mixture weights.

Furthermore, the list of central issues for future research clearly includes the topic of MSE estimation. The suggested approximation may prove to be a good starting point for improvements: It seems worth to further pursue the adopted strategy of approximating a suitably expressed MSE through known terms. It may, however, be necessary to also account for the uncertainty introduced through the estimation of the posterior probabilities of subgroup membership. Additionally, a suitable estimator has to be implemented for the mixture of unit levels, too.

If the suggested estimators are intended to be used in real-data applications, model selection and diagnostics are further important tasks for future research. While useful criteria for selecting the number of components could be provided, researchers also require guidelines on how to specify the component models. It is important to note that the suggested estimators generally allow for different sets of covariates in the K component models. This is clearly a potential strength of the approach, but adequate procedures have to be found to support the decision between component-specific or fixed sets of covariates and the subsequent selection of predictors. Improved procedures in this regard might also help to further enhance the results obtained in the application presented in this thesis.

Finally, it may prove interesting to investigate extensions to the suggested approach. Theoretically, mixtures of any kind of distributions are possible. More specifically, it might, for example, be interesting to estimate mixtures of a standard small area model and a spatial small area model.

Appendix

A.1 EM Algorithm

The EM algorithm introduced by Dempster et al. (1977) is an general-purpose numerical algorithm for calculating maximum likelihood estimates in the case of incomplete data or in cases that in a broad sense can be interpreted as a missing data situation in order simplify the estimation problem. It exploits the fact that maximum likelihood estimation would often be straightforward if the suitably defined complete (or augmented) data were observed and a complete-data log-likelihood could be formed. The algorithm imitates the simplified estimation problem by alternatively working on the expectation of the complete-data log-likelihood. See McLachlan and Krishnan (2008) for an comprehensive overview on the EM algorithm and its numerous extensions.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be an n-dimensional vector of observations from a random variable with density $f(\boldsymbol{y}|\boldsymbol{\psi})$. The likelihood function $L(\boldsymbol{\psi})$ is formed from the joint density by considering it as a function of the unknown parameters $\boldsymbol{\psi}$ for given realizations $\boldsymbol{y}$, i.e. $L(\boldsymbol{\psi}) = f(\boldsymbol{y}|\boldsymbol{\psi})$. The corresponding log-likelihood $\log(L(\boldsymbol{\psi}))$ is denoted as $l(\boldsymbol{\psi})$. Now, let there be a suitably defined complete-data vector $\boldsymbol{x}$, that contains both the observed values $\boldsymbol{y}$ and some additional data $\boldsymbol{z}$. Note that this might either be missing data in the classical sense or some well-chosen hypothetical information. Either way, the complete-data log-likelihood that could be formed if the complete data was observable is given by

$$l_c(\boldsymbol{\psi}) = \log(L_c(\boldsymbol{\psi})) = \log f_c(\boldsymbol{x}|\boldsymbol{\psi}).$$

To solve the incomplete-data log-likelihood, the EM algorithm proceeds by alternatively working on this complete data log-likelihood, alternating between two eponymous steps (McLachlan and Krishnan, 2008): The expectation step (*E-step*) and the maximization step (*M-step*). Generally, the following procedure is applied:

- *Specification of starting values*
  Choice of $\widehat{\boldsymbol{\psi}}^{(0)}$ as initial value for $\widehat{\boldsymbol{\psi}}$.

- Expectation-step *(E-step)*
  An expectation Q of the complete-data log-likelihood $l_c(\boldsymbol{\psi})$ is obtained, using the current estimates (or in the first step some initial value $\widehat{\boldsymbol{\psi}}^{(0)}$) of the model parameters $\boldsymbol{\psi}$. This particularly requires deriving the conditional expectation of the latent variable $\boldsymbol{z}$ given $\boldsymbol{y}$. It is

$$Q(\boldsymbol{\psi}; \widehat{\boldsymbol{\psi}}^{(t-1)}) = E_{\widehat{\boldsymbol{\psi}}^{(t)}}(l_c(\boldsymbol{\psi})|\boldsymbol{y}). \tag{A.1}$$

  Here and in the following $E_{\widehat{\boldsymbol{\psi}}^{(t-1)}}$ denotes expectation parametrized by $\widehat{\boldsymbol{\psi}}^{(t-1)}$, where $\widehat{\boldsymbol{\psi}}^{(t-1)}$ is the vector of parameter estimates obtained in the previous iteration step $(t-1)$.

- Maximization-step *(M-step)*,
  where an updated estimate $\widehat{\boldsymbol{\psi}}^{(t)}$ is obtained by maximizing Q with respect to $\boldsymbol{\psi}$ over the parameter space, i.e.

$$\widehat{\boldsymbol{\psi}}^{(t)} = \text{argmax}_{\boldsymbol{\psi}} Q(\boldsymbol{\psi}; \widehat{\boldsymbol{\psi}}^{(t-1)}). \tag{A.2}$$

  The solution often exists in closed form. For cases where global maximization of Q is still infeasible, DEMPSTER et al. (1977) suggested the generalized EM algorithm, which only requires a choice of $\widehat{\boldsymbol{\psi}}^{(t)}$ so that $Q(\widehat{\boldsymbol{\psi}}^{(t)}; \widehat{\boldsymbol{\psi}}^{(t-1)}) \geq Q(\widehat{\boldsymbol{\psi}}^{(t-1)}; \widehat{\boldsymbol{\psi}}^{(t-1)})$ is fulfilled (MCLACHLAN and KRISHNAN, 2008, Chapter 1.5.5).

- *Termination*:
  Both steps are repeated until the likelihood improvement in a step is smaller than an *ex ante* specified threshold ϵ, that is until $L(\boldsymbol{\psi}^{(t)}) - L(\boldsymbol{\psi}^{(t-1)}) < \epsilon$.

DEMPSTER et al. (1977) showed that the likelihood $L(\boldsymbol{\psi})$ is never decreased after an iteration step so that – for a sequence of likelihood values bounded above – convergence of the algorithm is guaranteed. For multimodal distributions this might, however, be convergence to a local maximum. The estimation result obtained then depends on the initial values. To overcome this issue, the algorithm is usually applied repeatedly with different starting values. For a detailed account of convergence properties of the EM algorithm see DEMPSTER et al. (1977), WU (1983) and MCLACHLAN and KRISHNAN (2008).

Compared to alternative iterative procedures such as the Fisher-scoring algorithm or the Newton-Raphson algorithm, the EM algorithm is relatively robust to the choice of initial values (Demidenko, 2004, Chapter 1.7). It is, however, known to converge slowly, i.e. to need a larger number of iteration steps than competing algorithms. The calculation time for an iteration is, however, usually low which offsets this disadvantage. An overview over methods proposed to speed up convergence is given by McLachlan and Krishnan (2008, Chapter 4 and 5). Another commonly stated drawback of the EM algorithm is that, different to both the Fisher-scoring and the Newton-Raphson algorithm, no asymptotic covariance matrix of the estimated parameters is obtained as a by-product of the estimation process. See McLachlan and Krishnan (2008, Chapter 4) for an extended review of methods to obtain the covariance matrix of ML estimates calculated using the EM algorithm.

A.2 Simulation Studies: Supplementary Material

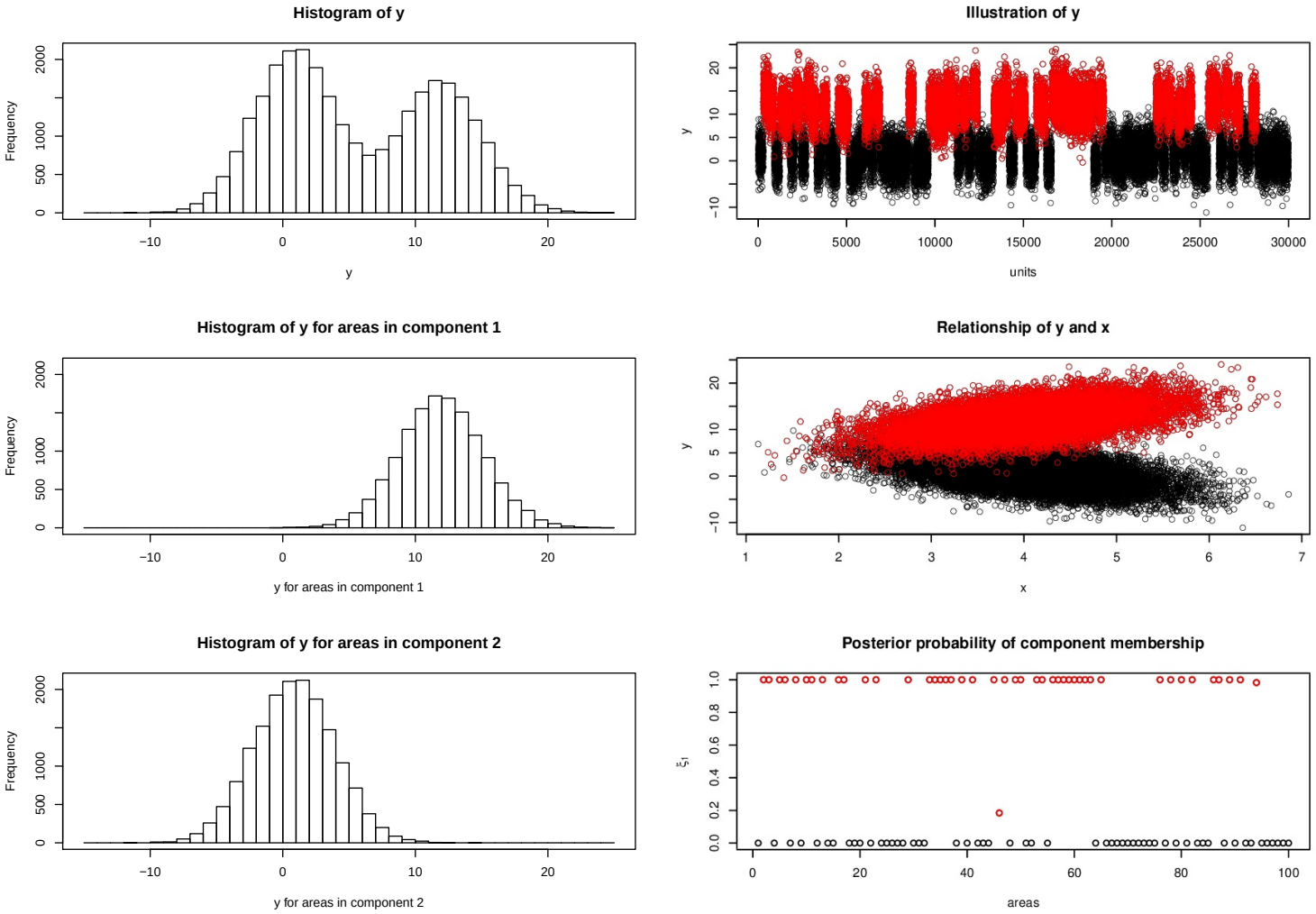

Figure A.1: Population 2

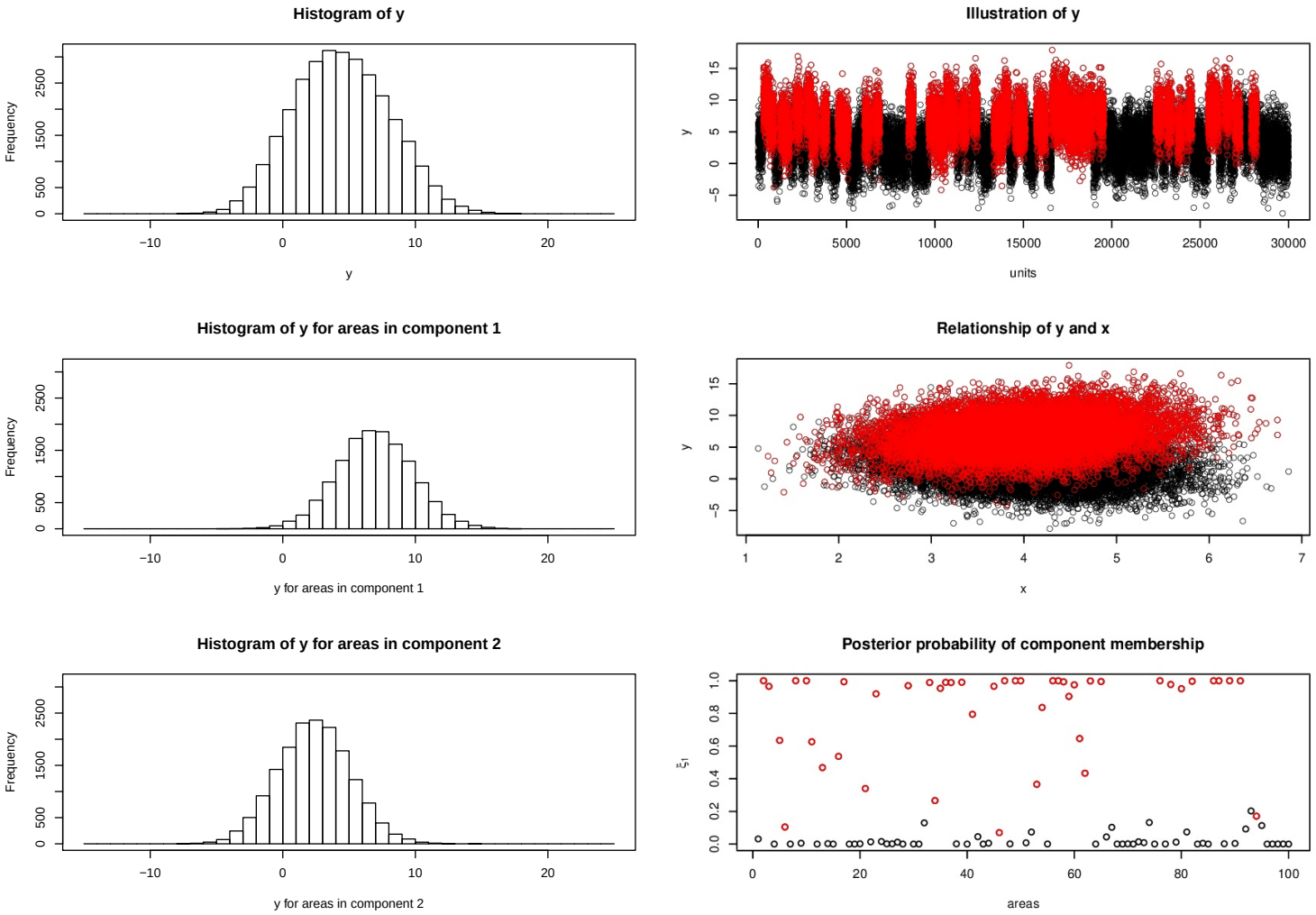

Figure A.2: Population 3

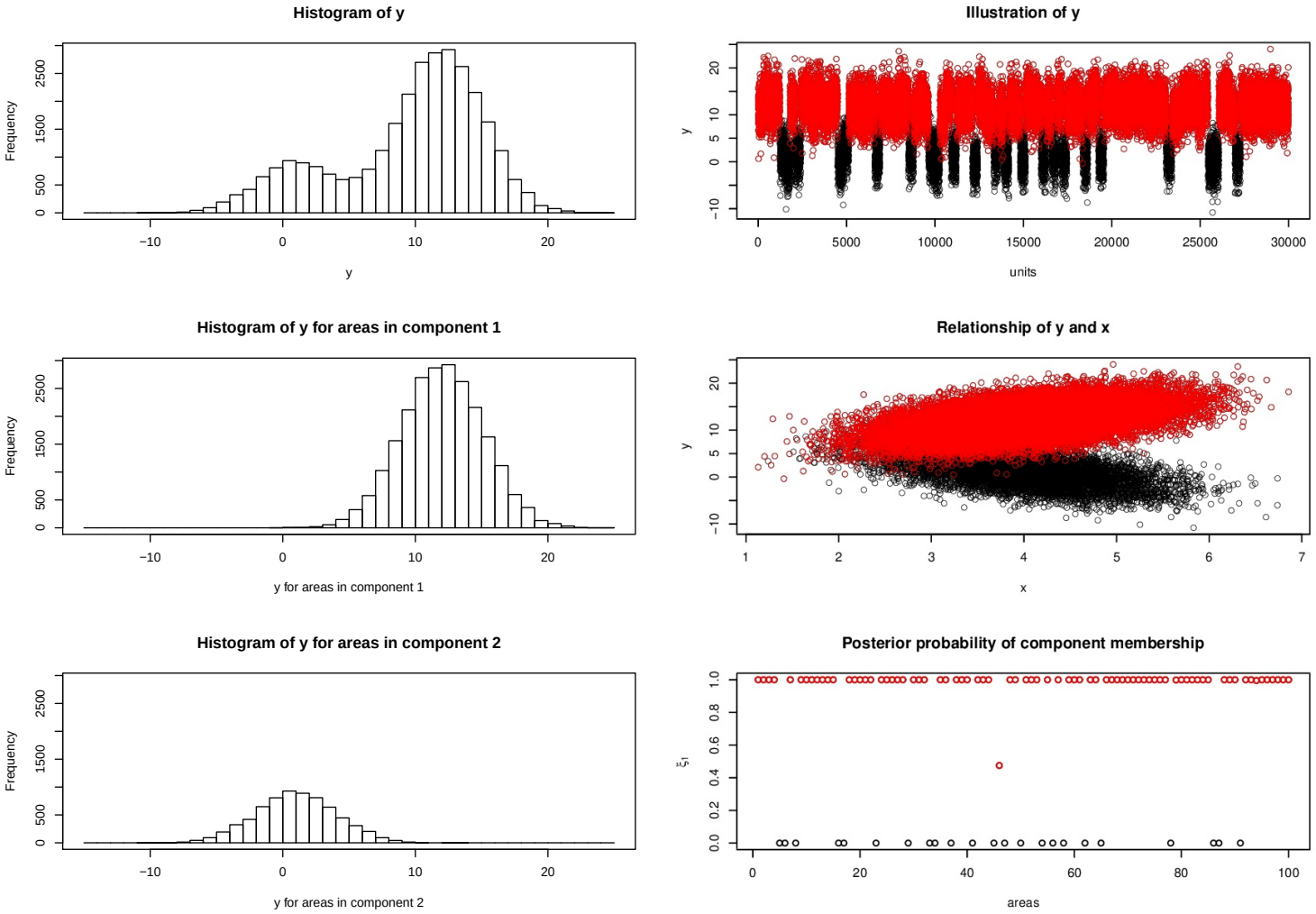

Figure A.3: Population 4

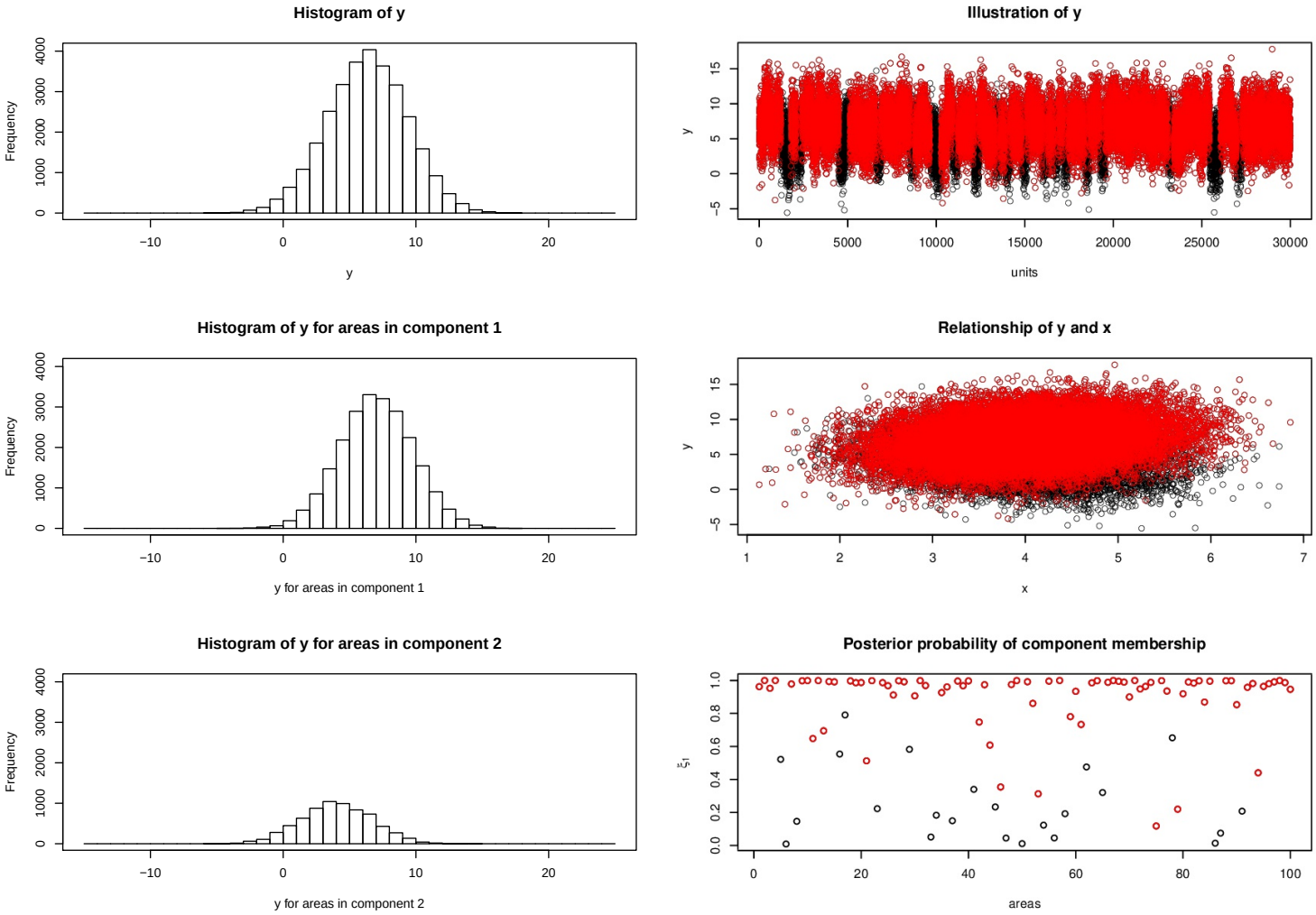

Figure A.4: Population 5

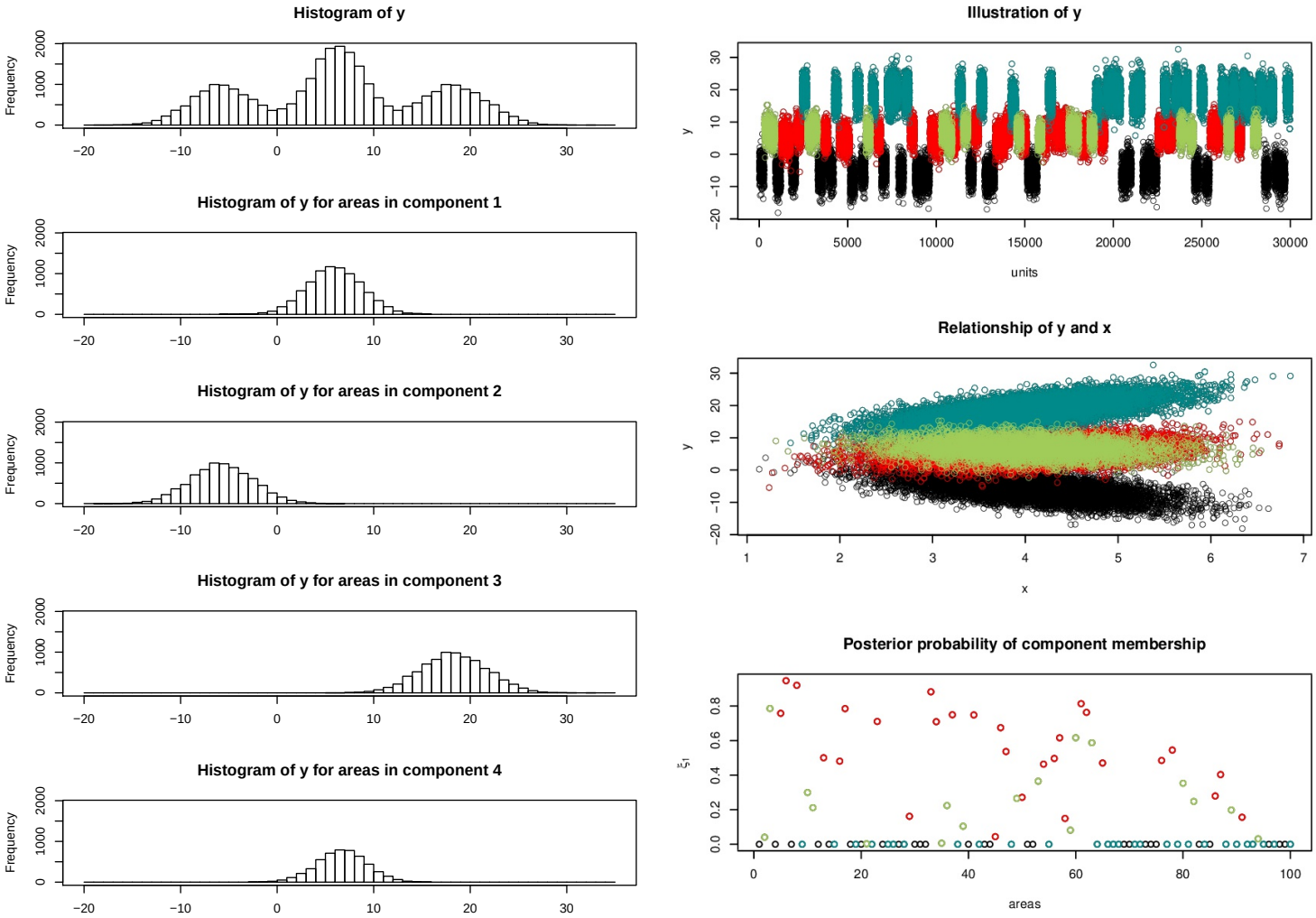

Figure A.5: Population 6

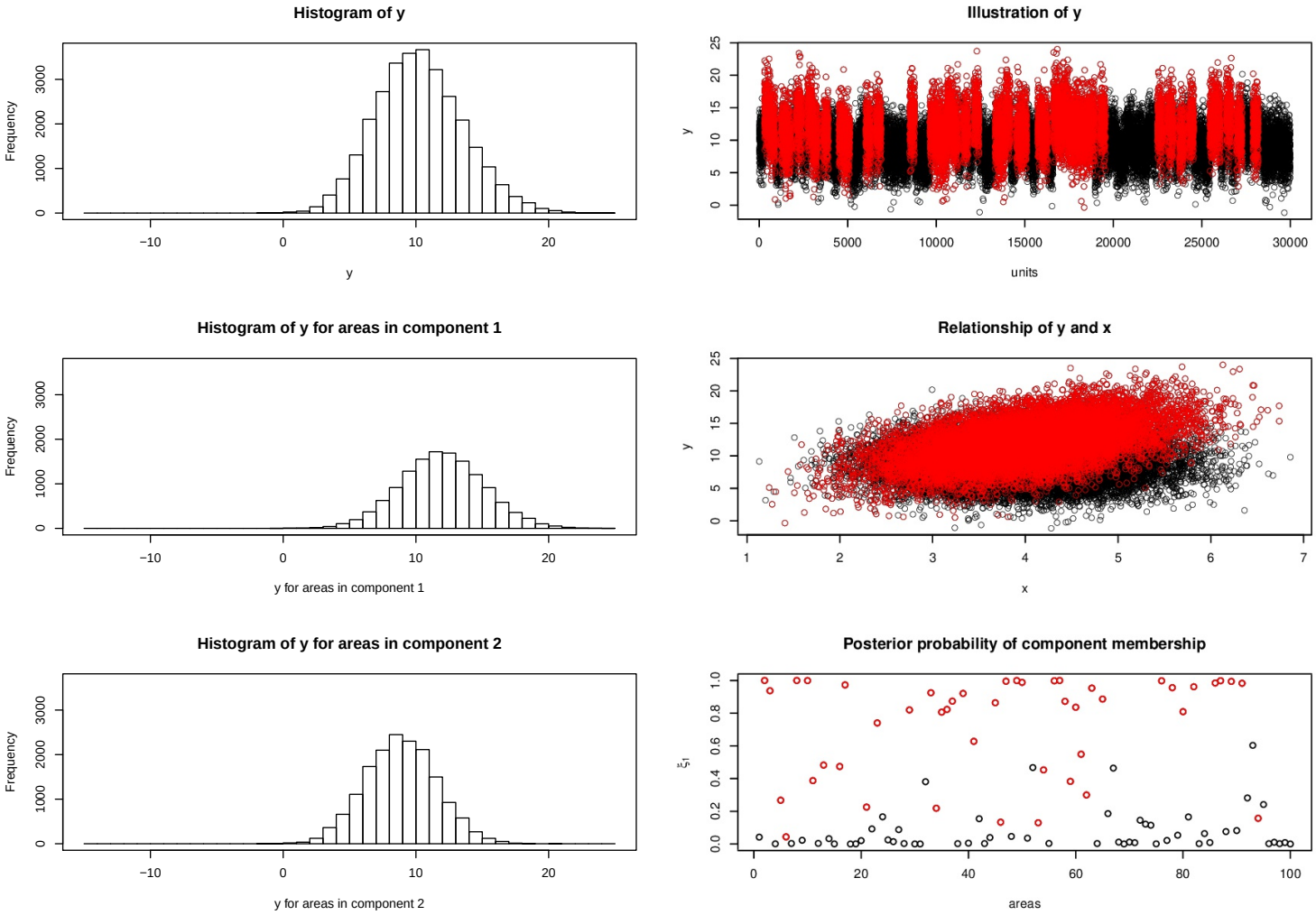

Figure A.6: Population 7

A.3 Application: Supplementary Material

A.3.1 Auxiliary Information

The regional indicators employed in the application mostly stem from the data collection INKAR (Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR), 2017), which is provided by German official statistics and openly accessible under `http://www.inkar.de`. Additionally, results from the German *Zensus* 2011 (Bayerisches Landesamt für Statistik, 2018), which can be retrieved from `https://ergebnisse.zensus2011.de`, are used. In what follows, brief descriptions of the indicators, their construction and the underlying data are given. If not stated otherwise, all information is based on the metadata provided with the data. Table A.1 contains descriptive statistics for all indicators.

Table A.1: Descriptive statistics for covariates

	Mean	SD	Median	Min	Max
PGRO	-2.32	3.00	-2.10	-9.40	6.30
RENT	48.73	13.59	45.85	24.10	82.50
VACQ	5.23	2.59	4.70	1.50	13.90
EMPL	50.53	3.85	50.60	37.90	61.00
MIGR	-1.59	4.50	-1.60	-11.20	22.10
LAND	94.13	77.92	70.65	4.80	497.20
DENS	1780.13	976.25	1471.55	512.40	5503.30

PGRO

The population growth rate (PGRO) measures the increase in the number of a district's residents between 2004 and 2009 in percent. It is based on the intercensal population updates annually provided by official statistics (for details on the methodology see Statistisches Bundesamt, 2008a).

RENT

The indicator prevalence of rented housing (RENT) measures the importance of rented housing in a district and is calculated as the ratio of dwellings rented out for residential purposes (including rent-free) to all inhabited and uninhabited dwellings in percent. It is based on data from the German *Zensus* 2011. See Bayerisches Landesamt für Statistik (2018) for details.

VACQ

The vacancy rate (VACQ) is calculated as the ratio of uninhabited dwellings to all inhabited and uninhabited dwellings in buildings with residential space in percent. It is based on data from the German *Zensus* 2011 (see Bayerisches Landesamt für Statistik (2018) for details).

EMPL

The employment rate (EMPL) is conceptualized as the ratio of employees subject to social insurance contributions with residence in the respective district to the district's working age population (aged 15 to 65 years). Note that a share of about 30% of the working population such as self-employed and civil servants are not included in this measure. The information is derived from the federal employment agencies register of all employees covered by social security (see Bundesagentur für Arbeit (2012)). The working age population is retrieved from the intercensal population updates described above.

MIGR

The net migration rate (MIGR) is the difference between the number of immigrants and the number of emigrants in a district, relative to the size of the district's resident population. It is reported per 1,000 residents over a period of one year. The information is retrieved from the migration statistics of the Federal Statistical Office, which is a register based on the registrations and deregistrations recorded by the registration offices (see Statistisches Bundesamt (2008b)).

LAND

The price of building land (LAND) measures the average price of building land per square meter. The indicator is based on the statistic on building land prices provided by official statistics in Germany. This is a secondary-statistical register based on information of the fiscal authorities and the *Gutachterausschüsse für Grundstückswerte* (independent expert panels institutionalized by federal law which monitor price development on the property market). See Statistisches Bundesamt (2010b) for details.

DENS

The settlement density (DENS) measures the per square kilometre of land under settlement and transport infrastructure (state of 31. December 2009). It is an adjusted form of the population density, where only area under residential use is considered. The number of inhabitants is obtained from the intercensal popula-

tion updates (see above). The area under settlement and transport infrastructure is available from the federal statistical office's register on actual land use (see Statistisches Bundesamt (2010a)).

A.3.2 Estimated Parameters

Table A.2: Estimated model parameters for competing estimators

	FH	FHS	FHclust		FHmix		FHmixconc		FHmix with DENS	
			k=1	k=2	k=1	k=2	k=1	k=2	k=1	k=2
$\hat{\sigma}_v^2$	0.081	0.039	0.028	0.002	0.033	0.006	0.025	0.007	0.036	1.75×10^{-7}
Intercept	2.823	2.758	2.566	3.44	2.169	3.472	2.126	3.59	1.891	3.036
PGRO	0.033	0.033	0.110	-3.25×10^{-5}	0.094	0.008	0.120	-0.004	0.118	1.619×10^{-2}
RENT	0.014	0.013	0.010	1.32×10^{-2}	0.013	0.013	0.009	0.012	-0.001	2.016×10^{-2}
VACQ	-0.039	-0.030	-0.096	-2.01×10^{-2}	-0.058	-0.017	-0.115	-0.015	-0.069	-2.70×10^{-3}
EMPL	0.024	0.026	0.044	8.52×10^{-3}	0.045	0.008	0.057	0.006	0.057	1.055×10^{-2}
MIGR	0.021	0.019	0.021	1.40×10^{-2}	0.035	0.009	0.001	0.029	0.040	6.24×10^{-3}
LAND	0.004	0.003	0.003	1.97×10^{-3}	0.004	0.002	0.003	0.002	1.57×10^{-4}	5.41×10^{-3}
DENS									3.94×10^{-4}	-2.14×10^{-4}
$\hat{\lambda}_k$			0.247	0.753	0.41	0.59	$\overline{\lambda}_{i,1} = 0.303$	$\overline{\lambda}_{i,2} = 0.697$	0.377	0.623

A.3.3 External Validation: Quoted Rents by the BBSR

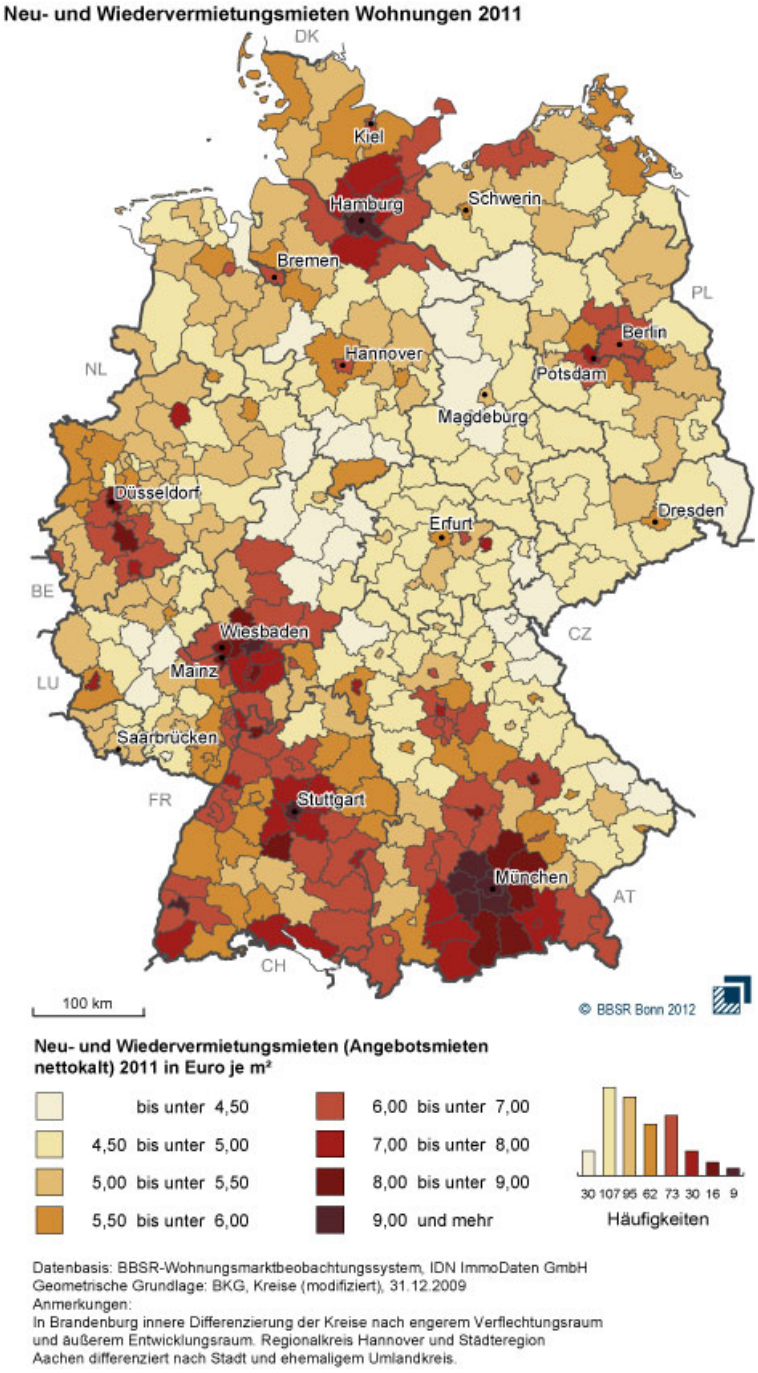

Figure A.7: Quoted rents by the BBSR

References

Ahmad, K. E. (**1988**): *Identifiability of finite mixtures using a new transform.* Annals of the Institute of Statistical Mathematics, 40 (2), pp. 261–265.

Aitkin, M. and **Rubin, D.** (**1985**): *Estimation and hypothesis testing in finite mixture models models.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 47, pp. 67–75.

Alfons, A., **Filzmoser, P.**, **Hulliger, B.**, **Kolb, J.-P.**, **Kraft, S.**, **Münnich, R.** and **Templ, M.** (**2011**): *Synthetic data generation of SILC data. AMELI Deliverable 6.2.* Technical report, University of Trier.
URL `https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Delivrables/AMELI-WP6-D6.2-240611.pdf`

Articus, C. I. (**2014**): *Small-Area-Verfahren zur Schätzung regionaler Mietpreise.* WISTA – Wirtschaft und Statistik, 2, pp. 113–118.

Articus, C. I. and **Burgard, J. P.** (**forthcoming**): *Finite Mixture Models for Small Area Estimation: The Case of Estimating Regional Rental Prices in Germany.* Under revision.

Bates, D., **Mächler, M.**, **Bolker, B.** and **Walker, S.** (**2015**): *Fitting Linear Mixed-Effects Models Using lme4.* Journal of Statistical Software, 67 (1), pp. 1–48.

Battese, G. E., **Harter, R. M.** and **Fuller, W. A.** (**1988**): *An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data.* Journal of the American Statistical Association, 83 (401), pp. 28 – 36.

Bauer, D. J. and **Curran, P. J.** (**2004**): *The Integration of Continuous and Discrete Latent Variable Models: Potential Problems and Promising Opportunities.* Psychological Methods, 9 (1), pp. 3–29.

Bayerisches Landesamt für Statistik (**2018**): *Census Database of the Census 2011 from Federal Statictical Offices.* `https://ergebnisse.zensus2011.de`, accessed: 2018-01-16.

Bell, W. R. (**2008**): *Examining Sensitivity of Small Area Inferences to Uncertainty About Sampling Error Variances.* Proceedings of the Section on Survey Research Methods, pp. 327–334, American Statistical Association.

Benaglia, T., **Chauveau, D.**, **Hunter, D. R.** and **Young, D.** (**2009**): *mixtools: An R Package for Analyzing Finite Mixture Models*. Journal of Statistical Software, 32 (6), pp. 1–29.

Biernacki, C., **Celeux, G.** and **Govaert, G.** (**1998**): *Assessing a mixture model for clustering with the integrated classification likelihood*. Technical report, INRIA.

Biernacki, C. and **Govaert, G.** (**1997**): *Using the classification likelihood to choose the number of clusters*. Computing Science and Statistics, 29 (2), pp. 451–457.

Binder, D. A. (**1978**): *Bayesian cluster analysis*. Biometrika, 65 (1), p. 31.

Booth, J., **Casella, G.** and **Hobert, J.** (**2008**): *Clustering using objective functions and stochastic search*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70, pp. 119–139.

Brown, G., **Chambers, R.**, **Heady, P.** and **Heasman, D.** (**2001**): *Evaluation of Small Area Estimation Methods – An Application to Unemployment Estimates from the UK LFS*. Proceedings of Statistics Canada Symposium 2001, Statistics Canada.

Bundesagentur für Arbeit (**2012**): Statistik der sozialversicherungspflichtigen und geringfügigen Beschäftigung. Nürnberg: Bundesagentur für Arbeit.

Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) (**2012**): *Mieten und Preise: Wohnimmobilien*. `http://www.bbsr.bund.de/BBSR/DE/WohnenImmobilien/Immobilienmarktbeobachtung/ProjekteFachbeitraege/MietenPreise/Mieten/Mieten.html`, accessed: 2018-06-01.

Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) (**2017**): *INKAR online*. `http://inkar.de/`, accessed: 2018-04-01.

Burgard, J. P. (**2013**): Evaluation of Small Area Techniques for Applications in Official Statistics. Ph.D. thesis, Universität Trier.

Burgard, J. P., **Kolb, J.-P.**, **Merkle, H.** and **Münnich, R.** (**2017**): *Synthetic data for open and reproducible methodological research in social sciences and official statistics*. AStA Wirtschafts- und Sozialstatistisches Archiv, 11 (3), pp. 233–244.

Burgard, J. P., **Münnich, R.** and **Zimmermann, T.** (**2016**): *Impact of Sampling Designs in Small Area Estimation with Applications to Poverty Measurement*. **Pratesi, M.** (editor) Analysis of Poverty Data by Small Area Estimation, pp. 83–108, John Wiley & Sons.

Burnham, K. P. and **Anderson, D. R.** (**2002**): Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer, 2 ed.

Celeux, G., **Martin, O.** and **Lavergne, C.** (**2005**): *Mixture of linear mixed*

models for clustering gene expression profiles from repeated microarray experiments. Statistical Modelling, 5, pp. 1–25.

Celeux, G. and **Soromenho, G.** (**1996**): *An entropy criterion for assessing the number of clusters in a mixture model*. Journal of Classification, 13 (2), pp. 195–212.

Chandra, H. and **Chambers, R.** (**2016**): *Small area estimation for semicontinuous data*. Biometrical Journal, 58 (2), ISSN 1521-4036.

Cole, V. T. and **Bauer, D. J.** (**2016**): *A Note on the Use of Mixture Models for Individual Prediction*. Structural Equation Modeling, 23 (4), pp. 615–631.

Crawford, S. (**1994**): *An application of the Laplace method to finite mixture distributions*. Journal of the American Statistical Association, 89, pp. 259–267.

Dang, U. J. and **McNicholas, P. D.** (**2015**): *Families of Parsimonious Finite Mixtures of Regression Models*. **Morlini, I.**, **Minerva, T.** and **Vichi, M.** (editors) Advances in Statistical Models for Data Analysis, pp. 73–84, Cham: Springer International Publishing.

Das, K., **Jiang, J.** and **Rao, J. N. K.** (**2004**): *Mean Squared Error of Empirical Predictor*. The Annals of Statistics, 32 (2), pp. 818–840.

Dasgupta, A. and **Raftery, A. E.** (**1998**): *Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering*. Journal of the American Statistical Association, 93 (441), pp. 294–302.

Datta, G. S. and **Lahiri, P.** (**1995**): *Robust Hierarchical Bayes Estimation of Small Area Characteristics in the Presence of Covariates and Outliers*. Journal of Multivariate Analysis, 54, pp. 310–328.

Datta, G. S. and **Lahiri, P.** (**2000**): *A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems*. Statistica Sinica, 10, pp. 613–627.

Datta, G. S. and **Mandal, A.** (**2015**): *Small Area Estimation with Uncertain Random Effects*. Journal of the American Statistical Association, 110 (512), pp. 1735–1744.

Datta, G. S., **Rao, J. N. K.** and **Smith, D. D.** (**2005**): *On measuring the variability of small area estimators under a basic area level model*. Biometrika, 92 (1), pp. 183–196.

Day, N. E. (**1969**): *Estimating the components of a mixture of two normal distributions*. Biometrika, 56 (3), pp. 463–474.

Dayton, C. M. and **Macready, G. B.** (**1988**): *Concomitant-Variable Latent-Class Models*. Journal of the American Statistical Association, 83 (401), pp. 173–178.

de Leeuw, J. and **Meijer, E.** (**2008**): *Introduction to Multilevel Analysis*. **de Leeuw, J.** and **Meijer, E.** (editors) Handbook of Multilevel Analysis,

pp. 1–75, New York: Springer.

Demidenko, E. (**2004**): Mixed Models. Wiley series in probability and statistics, Hoboken: John Wiley & Sons.

Dempster, A. P., **Laird, N. M.** and **Rubin, D. B.** (**1977**): *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39 (1), pp. 1–38.

DeSarbo, W. S. and **Cron, W. L.** (**1988**): *A maximum likelihood methodology for clusterwise linear regression.* Journal of Classification, 5 (2), pp. 249–282.

Du, Y., **Kahili, A.**, **Neslehova, J. G.** and **Steele, R. J.** (**2013**): *Simultaneous fixed and random effects selection in finite mixture of linear mixed-effects models.* Canadian Journal of Statistics, 41 (4), pp. 596–616.

Elbers, C. and **van der Weide, R.** (**2014**): *Estimation of Normal Mixtures in a Nested Error Model with an Application to Small Area Estimation of Poverty and Inequality.* Technical report, World Bank Group.

Everitt, B. S. and **Hand, D. J.** (**1981**): Finite Mixture Distributions. London: Chapman & Hall.

Fabrizi, E., **Montanari, G. E.** and **Ranalli, M. G.** (**2016**): *A hierarchical latent class model for predicting disability small area counts from survey data.* Journal of the Royal Statistical Society: Series A (Statistics in Society), 179 (1), pp. 103–131.

Fahrmeir, L., **Kneib, T.**, **Lang, S.** and **Marx, B.** (**2013**): Regression. Models, Methods and Applications. Heidelberg et al.: Springer.

Fair, R. C. and **Jaffee, D. M.** (**1972**): *Methods of estimation for markets in disequilibrium.* Econometrica, 40 (3), pp. 497–514.

Farewell, V. T. (**1982**): *The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors.* Biometrics, 38 (4), pp. 1041–1046.

Fay, R. E. and **Herriot, R. A.** (**1979**): *Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data.* Journal of the American Statistical Association, 74 (366), pp. 269–277.

Feller, W. (**1943**): *On a General Class of "Contagious" Distributions.* The Annals of Mathematical Statistics, 14 (4), pp. 389–400.

Fonseca, J. R. S. and **Cardoso, M. G. M. S.** (**2007**): *Mixture-model Cluster Analysis Using Information Theoretical Criteria.* Intelligent Data Analysis, 11 (2), pp. 155–173.

Fraley, C. and **Raftery, A. E.** (**2002**): *Model-Based Clustering, Discriminant Analysis, and Density Estimation.* Journal of the American Statistical Association, 97 (458), pp. 611–631.

Frühwirth-Schnatter, S. (**2006**): Finite Mixture and Markov Switching Models. Springer Series in Statistics, New York: Springer.

Gershunskaya, J. (**2010**): *Robust Small Area Estimation Using a Mixture Model.*

Proceedings of the Section on Survey Research Methods, pp. 2783–2796, American Statistical Association.

Ghosh, M. and **Rao, J. N. K.** (**1994**): *Small Area Estimation: An Appraisal*. Statistical Science, 9 (1), pp. 55–93.

Goldberger, A. S. (**1962**): *Best Linear Unbiased Prediction in the Generalized Linear Regression Model*. Journal of the American Statistical Association, 57 (298), pp. 369–375.

Gormley, I. C. and **Murphy, T. B.** (**2011**): *Mixture of experts modelling with social science applications*. **Mengersen, K. L.**, **Robert, C. P.** and **Titterington, D. M.** (editors) Mixtures: Estimation and Applications, Wiley series in probability and statistics, pp. 101–121, New York: John Wiley & Sons.

Grilli, L., **Rampichini, C.** and **Varriale, R.** (**2015**): *Binomial Mixture Modeling of University Credits*. Communications in Statistics - Theory and Methods, 44 (22), pp. 4866–4879.

Grün, B. (**2008**): *Fitting finite mixtures of linear mixed models with the EM algorithm*. **Brito, P.** (editor) Compstat 2008 – Proceedings in Computational Statistics, pp. 165–173, Physica Verlag.

Grün, B. and **Leisch, F.** (**2007**): *Fitting Finite Mixtures of Generalized Linear Regressions in R*. Computational Statistics and Data Analysis, 51 (11), pp. 5247–5252.

Grün, B. and **Leisch, F.** (**2008**): *Finite Mixtures of Generalized Linear Regression Models*. **Shalab** and **Heumann, C.** (editors) Recent Advances in Linear Models and Related Areas: Essays in Honour of Helge Toutenburg, pp. 205–230, Heidelberg: Physica-Verlag HD.

Gudicha, D. W. and **Vermunt, J. K.** (**2013**): *Mixture Model Clustering with Covariates Using Adjusted Three-Step Approaches*. **Lausen, B.**, **Van den Poel, D.** and **Ultsch, A.** (editors) Algorithms from and for Nature and Life: Classification and Data Analysis, pp. 87–94, Cham: Springer International Publishing.

Hartley, H. O. and **Rao, J. N. K.** (**1967**): *Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model*. Biometrika, 54 (1/2), pp. 93–108.

Harville, D. A. (**1974**): *optimal procedures for some constrained selection problems*. Journal of the American Statistical Association, 69, pp. 446–452.

Harville, D. A. (**1976**): *Extension of the Gauss-Markov Theorem to include the Estimation of Random Effects*. The Annals of Statistics, 4 (2), pp. 384–395.

Harville, D. A. and **Jeske, D. R.** (**1992**): *Mean Squared Error of Estimation or Prediction Under a General Linear Model*. Journal of the American Statistical Association, 87 (419), pp. 724–731.

Hawkins, D. S., **Allen, D. M.** and **Stromberg, A. J.** (**2001**): *Determining the*

number of components in mixtures of linear models. Computational Statisics and Data Analysis, 38 (1), pp. 14–48.

Henderson, C. R. (**1950**): *Estimation of genetic parameters (abstract)*. Annals of Mathematical Statistics, 21, pp. 309–310.

Henderson, C. R. (**1963**): *Selection Index and Expected Genetic Advance*. Statistical Genetics and Plant Breeding, pp. 141–163.

Henderson, C. R. (**1973**): *Sire evaluation and genetic trends*. Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush, pp. 10–41, Champaign, IL: American Society of Animal Science and American Dairy Science Association.

Henderson, C. R. (**1975**): *Best Linear Unbiased Estimation and Prediction under a Selection Model*. Biometrics, 31 (2), pp. 423–447.

Henderson, C. R., **Kempthorne, O.**, **Searle, S. R.** and **von Krosigk, C. M.** (**1959**): *The Estimation of Environmental and Genetic Trends from Records Subject to Culling*. Biometrics, 15 (2), pp. 192–218.

Hennig, C. (**2000**): *Identifiablity of Models for Clusterwise Linear Regression*. Journal of Classification, 17 (2), pp. 273–296.

Henningsen, A. and **Toomet, O.** (**2011**): *maxLik: A package for maximum likelihood estimation in R*. Computational Statistics, 26 (3), pp. 443–458.

Hettmansperger, T. P. and **Thomas, H.** (**2000**): *Almost nonparametric inference for repeated measures in mixture models*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62 (4), pp. 811–825.

Holzmann, H., **Munk, A.** and **Gneiting, T.** (**2006**): *Identifiability of Finite Mixtures of Elliptical Distributions*. Scandinavian Journal of Statistics, 33 (4), pp. 753–763.

Hosmer, D. W. (**1974**): *Maximum likelihood estimates of the parameters of a mixture of two regression lines*. Communications in Statistics, Part A – Theory and Methods, 3, pp. 995–1006.

Jacobs, R. A., **Jordan, M. I.**, **Nowlan, S. J.** and **Hinton, G. E.** (**1991**): *Adaptive Mixtures of Local Experts*. Neural Computation, 3 (1), pp. 79–87.

Jennrich, R. I. and **Schluchter, M. D.** (**1986**): *Unbalanced Repeated-Measures Models with Structured Covariance Matrices*. Biometrics, 42 (4), pp. 805–820.

Jiang, J. (**2007**): Linear and Generalized Linear Mixed Models and their Applications. New York: Springer.

Jiang, J. (**2017**): Asymptotic Analysis of Mixed Effects Models: Theory, Applications, and Open Problems. Boca Raton: Chapman & Hall.

Jiang, J. and **Lahiri, P.** (**2006**): *Mixed model prediction and small area estimation*. TEST: An Official Journal of the Spanish Society of Statistics and Operations Research, 15 (1), pp. 1–96.

Jiang, W. and **Tanner, M. A.** (**1999**): *Om the identifiability of mixtures of*

experts. Neural Networks, 12, pp. 1253–1258.

Kackar, R. N. and **Harville, D. A.** (**1981**): *Unbiasedness of two-stage estimation and prediction procedures for mixed linear models*. Communications in Statistics. Series A, 10 (13), pp. 1249–1261.

Kackar, R. N. and **Harville, D. A.** (**1984**): *Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models*. Journal of the American Statistical Association, 79 (388), pp. 853–862.

Keribin, K. (**2000**): *Consistent Estimation of the Order of Mixture Models*. Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 62 (1), pp. 49–66.

Kopsch, A. (**2001**): Marktabgrenzung: Ein simultaner produkt- und nachfragerbezogener Ansatz. Wiesbaden: Springer.

Laird, N. M., **Lange, N.** and **Stram, D.** (**1987**): *Maximum likelihood computations with repeated measures: application of the EM algorithm*. Journal of the American Statistical Association, 82 (397), pp. 97–105.

Laird, N. M. and **Ware, J. J.** (**1982**): *Random-effects models for longitudinal data*. Biometrics, 38, pp. 963–974.

Leeflang, P., **Wittink, D.**, **Wedel, M.** and **Naert, P.** (**2000**): Building Models for Marketing Decisions. Boston et al.: Springer.

Leisch, F. (**2004**): *FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R*. Journal of Statistical Software, 11 (8), pp. 1–18.

Lenk, P. J. and **DeSarbo, W. S.** (**2000**): *Bayesian inference for finite mixtures of generalized linear models with random effects*. Psychometrika, 65 (1), pp. 93–119.

Leroux, B. G. (**1992**): *Consistent estimation of a mixing distribution*. The Annals of Statistics, 20 (1), pp. 1350–1360.

Lindsay, B. G. (**1995**): Mixture Models: Theory, Geometry and Applications. NSF-CMBS Regional Conference Series in Probability and Statistics, Hayward: Institute of Mathematical Statistics.

Lindstrom, M. J. and **Bates, D. M.** (**1988**): *Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data*. Journal of the American Statistical Association, 83 (404), pp. 1014–1022.

Maiti, T. (**2003**): *Modelling Small Area Effects using Mixture of Gaussians*. The Indian Journal of Statistics, 65, pp. 612–625.

Maiti, T., **Ren, H.**, **Dass, S. C.**, **Lim, C.** and **Maier, K. S.** (**2014**): *Clustering-Based Small Area Estimation: An Application to MEAP Data*. Calcutta Statistical Association Bulletin, 66, pp. 73–93.

Marron, J. S. and **Wand, M. P.** (**1992**): *Exact Mean Integrated Squared Error*. The Annals of Statistics, 20 (2), pp. 712–736.

Martella, F., **Vermunt, J. K.**, **Beekman, M.**, **Westendorp, R. G. J.**, **Slagboom, P. E.** and **Houwing-Duistermaat, J. J.** (**2011**): *A mixture model with random-effects components for classifying sibling pairs*. Statistics in Medicine, 30 (27), pp. 3252–3264.

Martinez, M. J., **Lavergne, C.** and **Trottier, C.** (**2009**): *A mixture model-based approach to the clustering of exponential repated data*. Journal of Multivariate Analysis, 100, pp. 1938–1951.

McCulloch, C. E., **Searle, S. R.** and **Neuhaus, J. M.** (**2008**): Generalized, Linear, and Mixed Models. Wiley series in probability and statistics, New York: John Wiley & Sons, 2 ed.

McLachlan, G. J. and **Basford, K. E.** (**1988**): Mixture Models: Inference and Applications to Clustering. New York/Basel: Marcel Dekker.

McLachlan, G. J. and **Krishnan, T.** (**2008**): The EM algorithm and extensions. Hoboken: Wiley, 2 ed.

McLachlan, G. J., **Ng, S. K.** and **Wang, K.** (**2008**): *Clustering via Mixture Regression Models with Random Effects*. COMPSTAT. Proceedings in Computational Statistics, Physica-Verlag.

McLachlan, G. J. and **Peel, D.** (**2000**): Finite Mixture Models. Wiley series in probability and statistics, New York: John Wiley & Sons.

McLachlan, G. J. and **Rathnayake, S.** (**2014**): *On the number of components in a Gaussian mixture model*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4 (5), pp. 341–355.

McLachlan, N. G. and **Ng, S. K.** (**2000**): *A comparison of some information criteria for the number of components in a mixture model*. Technical report, Brisbane: Department of Mathematics.

Münnich, R. (**2014**): *Small area applications: some remarks from a design-based view*. Presentation at the SAE2014 conference, Poznan 2014. `http://sae2014.ue.poznan.pl/presentations/SAE2014_Ralf_Munnich_c330a31c0a.pdf`.

Münnich, R., **Schürle, J.**, **Bihler, W.**, **Boonstra, H.-J.**, **Knottnerus, P.**, **Nieuwenbroek, N.**, **Haslinger, A.**, **Laaksonen, S.**, **Wiegert, R.**, **Eckmair, D.**, **Quatember, A.**, **Wagner, H.**, **Renfer, J.-P.** and **Oetliker, U.** (**2003**): *Monte Carlo simulation study of European surveys – DACSEIS deliverables 3.1 and 3.2*. Technical report, University of Tübingen.
URL `https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Dacseis_Deliverables/DACSEIS-D3-1-D3-2.pdf`

Molina, I. and **Marhuenda, Y.** (**2015**): *sae: An R Package for Small Area Estimation*. The R Journal, 7 (1), pp. 81–98.
URL `http://journal.r-project.org/archive/2015-1/molina-marhuenda.pdf`

Molina, I. and **Rao, J. N. K.** (**2010**): *Small area estimation of poverty indicators*. Canadian Journal of Statistics, 38 (3), pp. 369–385.

Molina, I., **Salvati, N.** and **Pratesi, M.** (**2009**): *Bootstrap for estimating the MSE of the Spatial EBLUP*. Computational Statistics, 24 (3), pp. 441–458.

Münnich, R. T., **Burgard, J. P.** and **Vogt, M.** (**2013**): *Small Area-Statistik: Methoden und Anwendungen*. AStA Wirtschafts- und Sozialstatistisches Archiv, 6 (3), pp. 149–191.

Nagin, D. S. (**2005**): Group-Based Modeling and Development. Boston: Harvard University Press.

Ng, S. K. and **McLachlan, G. J.** (**2014**): *Mixture models for clustering multilevel growth trajectories*. Computational Statistics and Data Analysis, 71, pp. 43–51.

Ng, S. K., **McLachlan, G. J.**, **Wang, K.**, **Ben-Tovim Jones, L.** and **Ng, S.-W.** (**2006**): *A mixture model with random-effects components for clustering correlated gene-expression profiles*. Bioinformatics, 22 (14), pp. 1745–1752.

Patterson, H. and **Thompson, R.** (**1971**): *Recovery of inter-block information when block sizes are unequal*. Biometrika, 58, pp. 545–554.

Peng, F., **Jacobs, R. A.** and **Tanner, M. A.** (**1996**): *Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models With an Application to Speech Recognition*. Journal of the American Statistical Association, 91 (435), pp. 953–960.

Pereira, L. N. and **Coelho, P. S.** (**2013**): *Estimation of house prices in regions with small sample sizes*. The Annals of Regional Science, 50 (2), pp. 603–621.

Pfeffermann, D. (**2002**): *Small Area Estimation: New Developments and Directions*. International Statistical Review / Revue Internationale de Statistique, 70 (1), pp. 125–143.

Pfeffermann, D. (**2013**): *New Important Developments in Small Area Estimation*. Statistical Science, 28 (1), pp. 40–68.

Pinheiro, J., **Bates, D.**, **DebRoy, S.**, **Sarkar, D.** and **R Core Team** (**2017**): nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131.
URL `https://CRAN.R-project.org/package=nlme`

Prasad, N. G. N. and **Rao, J. N. K.** (**1990**): *The Estimation of the Mean Squared Error of Small Area Estimators*. Journal of the American Statistical Association, 85 (409), pp. 163–171.

Pratesi, M. and **Salvati, N.** (**2008**): *Small area estimation: the EBLUP estimator based on spatially correlated random area effects*. Statistical Methods and Applications, 17 (1), pp. 113–141.

Quandt, R. E. (**1972**): *A new approach to estimating switching regressions*. Journal of the American Statistical Association, 67 (338), pp. 306–310.

Quandt, R. E. and **Ramsey, J. B.** (**1978**): *Estimating mixtures of normal distributions and switching regressions*. Journal of the American Statistical Association, 73 (364), pp. 730–752.

Rao, J. N. K. (**2003**): Small Area Estimation. Wiley series in survey methodology, New York: John Wiley & Sons.

Rao, J. N. K. and **Molina, I.** (**2015**): Small Area Estimation. Wiley series in survey methodology, New York: John Wiley & Sons, 2 ed.

Redner, R. A. and **Walker, H.** (**1984**): *Mixture densities, maximum likelihood and the EM algorithm*. SIAM Review, 26, pp. 195–239.

Ren, H. (**2011**): Some new models for small area estimation. Ph.D. thesis, Michigan State University.

Robinson, G. K. (**1991**): *That BLUP is a Good Thing: The Estimation of Random Effects*. Statistical Science, 6 (1), pp. 15–32.

Roeder, K. and **Wasserman, L.** (**1997**): *Practical Bayesian Density Estimation Using Mixtures of Normals*. Journal of the American Statistical Association, 89 (439), pp. 894–902.

Salvati, N., **Chandra, H.**, **Ranalli, M. G.** and **Chambers, R.** (**2010**): *Small area estimation using a nonparametric model-based direct estimator*. Computational Statistics and Data Analysis, 54 (9), pp. 2159 – 2171.

Sarstedt, M. and **Schwaiger, M.** (**2008**): *Model Selection in Mixture Regression Analysis–A Monte Carlo Simulation Study*. **Preisach, C.**, **Burkhardt, H.**, **Schmidt-Thieme, L.** and **Decker, R.** (editors) Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007, pp. 61–68, Berlin, Heidelberg: Springer Berlin Heidelberg.

Scharl, T., **Grün, B.** and **Leisch, F.** (**2010**): *Modelling time course gene expression data with finite mixtures of linear additive models*. Bioinformatics, 26 (3), pp. 370–377.

Schmid, T., **Tzavidis, N.**, **Münnich, R.** and **Chambers, R.** (**2016**): *Outlier Robust Small-Area Estimation Under Spatial Correlation*. Scandinavian Journal of Statistics, 43 (3), pp. 806–826.

Schwarz, G. (**1978**): *Estimating the dimension of a model*. Annals of Statistics, 6 (2), pp. 461–464.

Sclove, S. L. (**1987**): *Application of model-selection criteria to some problems in multivariate analysis*. Psychometrika, 52 (3), pp. 333–343.

Searle, S. R., **Casella, G.** and **McCulloch, C. E.** (**1992**): Variance Components. Wiley series in probability and statistics, New York: John Wiley & Sons.

Skrondal, A. and **Rabe-Hesketh, S.** (**2009**): *Prediction in multilevel generalized linear models*. Journal of the Royal Statistical Society: Series A (Statis-

tics in Society), 172 (3), pp. 659–687.

Statistisches Bundesamt (**2008**a): Fortschreibung des Bevölkerungsstandes. Qualitätsbericht. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (**2008**b): Wanderungsstatistik. Qualitätsbericht. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (**2010**a): Bodenfläche nach Art der tatsächlichen Nutzung. Qualitätsbericht. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (**2010**b): Preise. Kaufwerte für Bauland. Fachserie 17. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (**2011**): Mikrozensus 2010. Qualitätsbericht. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (**2012**): Bauen und Wohnen. Mikrozensus-Zusatzerhebung 2010. Wiesbaden: Statistisches Bundesamt.

Teicher, H. (**1961**): *Identifiability of Mixtures*. The Annals of Mathematical Statistics, 32 (1), pp. 244–248.

Teicher, H. (**1963**): *Identifiability of Finite Mixtures*. The Annals of Mathematical Statistics, 34 (4), pp. 1265–1269.

Thompson, T. J., **Smith, P. J.** and **Boyle, J. P.** (**1998**): *Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes*. Journal of the Royal Statistical Society: Series C (Applied Statistics), 47 (3), pp. 393–404.

Titterington, D. M., **Smith, A. F. M.** and **Makov, E.** (**1985**): Statistical Analysis of Finite Mixture Distributions. Chichester: John Wiley & Sons.

Torkashvand, E., **Jafari Jozani, M.** and **Torabi, M.** (**2017**): *Clustering in small area estimation with area level linear mixed models*. Journal of the Royal Statistical Society: Series A (Statistics in Society), p. to appear, ISSN 1467-985X.
URL `http://dx.doi.org/10.1111/rssa.12308`

Vaida, F. and **Blanchard, S.** (**2005**): *Conditional Akaike information for mixed-effects models*. Biometrika, 92 (2), pp. 351–370.

Venables, W. N. and **Ripley, B. D.** (**2002**): Modern Applied Statistics with S. New York: Springer, fourth ed.

Verbeke, G. and **Lesaffre, E.** (**1996**): *A Linear Mixed-Effects Model with Heterogeneity in the Random Effects Population*. Journal of the American Statistical Association, 91 (433), pp. 217–221.

Verbeke, G. and **Molenberghs, G.** (**2000**): Linear Mixed Models for Longitudinal Data. New York: Springer.

Wagner, J., **Münnich, R.**, **Hill, J.**, **Stoffels, J.** and **Udelhoven, T.** (**2017**): *Non-parametric small area models using shape-constrained penalized B-splines*. Journal of the Royal Statistical Society: Series A (Statistics

in Society), 180 (4), pp. 1089–1109.

Wang, K., **Ng, S. K.** and **McLachlan, G. J.** (**2012**): *Clustering of time-course gene expression profiles using normal mixture models with autoregressive random effects*. BMC Bioinformatics, 300 (13).

Wasserman, L. (**2012, August 4**): *Mixture Models: The Twilight Zone of Statistics (blog post)*. Retrieved from: `https://normaldeviate.wordpress.com/2012/08/04/mixture-models-the-twilight-zone-of-statistics/` (2017, March 14).

Wedel, M. (**2002**): *Concomitant variables in finite mixture models*. Statistica Neerlandica, 56 (3), pp. 362–375.

Wedel, M. and **Desarbo, M. S.** (**2002**): *Mixture Regression Models*. Applied Latent Class Analysis, pp. 366–382, Cambridge: Cambridge University Press.

Wedel, M. and **Kamakura, W. A.** (**2000**): Market Segmentation. Conceptual and Methodological Foundations. Boston: Springer, 2 ed.

Windham, M. P. and **Cutler, A.** (**1992**): *Information Ratios for Validating Mixture Analyses*. Journal of the American Statistical Association, 87 (420), pp. 1188–1192.

Wu, C. F. J. (**1983**): *On the convergence properties of the EM algorithm*. The Annals of Statistics, 11 (1), pp. 95–103.

Xu, W. and **Hedeker, D.** (**2001**): *A random-effects mixture model for classifying treatment response in longitudinal clinical trials*. Journal of biopharmaceutical statistics, 11 (4), pp. 253–273.

Yakowitz, S. and **Spragins, J. D.** (**1968**): *On the identifiability of finite mixtures*. Annals of Mathematical Statistics, 39, pp. 209–214.

Yang, C. C. and **Yang, C. C.** (**2007**): *Separating Latent Classes by Information Criteria*. Journal of Classification, 24 (2), pp. 183–203.

Yau, K. K. W., **Lee, A. H.** and **Ng, S. K.** (**2003**): *Finite Mixture Regression model with random effects: Application to neonatal hospital length of stay*. Computational Statistics and Data Analysis, 41, pp. 359–366.

Yuksel, S. E., **Wilson, J. N.** and **Gader, P. D.** (**2012**): *Twenty years of mixture of experts*. IEEE Transactions on Neural Networks and Learning Systems, 23, pp. 1177–1193.

Zimmermann, T. (**2015**): The interplay between sampling design and statistical modelling in small area estimation. Ph.D. thesis, Trier University.