Thesis submitted in fulfilment of the requirements for the degree

Dr. rer. pol.

on the topic

Recent Advances in Small Area Estimation of Economic and Poverty Indicators using Traditional and Alternative Data

to the

Chair of Statistics and Econometrics
Faculty of Social Sciences, Economics, and Business Administration
University of Bamberg

submitted by

Yeonjoo Lee
born in Seoul, Republic of Korea

Bamberg 2024

Cumulative dissertation
Yeonjoo Lee, *Recent Advances in Small Area Estimation of Economic and Poverty Indicators using Traditional and Alternative Data*

This thesis has been submitted to the Faculty of Social Sciences, Economics and Business Administration at the University of Bamberg as a dissertation.

First reviewer: Prof. Dr. Timo Schmid
Second reviewer: Prof. Dr. Natalia Rojas-Perilla
Date of defense: 24.07.2024

Diese Arbeit hat der Fakultät Sozial- und Wirtschaftswissenschaften der Otto-Friedrich-Universität Bamberg als Dissertation vorgelegen.

Erstgutachter: Prof. Dr. Timo Schmid
Zweitgutachterin: Prof. Dr. Natalia Rojas-Perilla
Tag der mündlichen Prüfung: 24.07.2024

Dieses Werk ist als freie Onlineversion über das Forschungsinformationssystem (FIS; https://fis.uni-bamberg.de) der Universität Bamberg erreichbar.

Das Werk steht unter der CC-Lizenz CC BY.

Lizenzvertrag: Creative Commons Namensnennung 4.0

https://creativecommons.org/licenses/by/4.0/

URN: urn:nbn:de:bvb:473-irb-1030636

DOI: https://doi.org/10.20378/irb-103063

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Dr. Timo Schmid. Thanks to his support, input and trust, I learned what studying, teaching, and academic research actually mean. Thanks to him, I could produce a dissertation that turned out better than what I could have hoped for. Ever since joining his team as a student it was not only his professional but also his personal support and encouragement that made it possible for me to be a part of the scientific community.

I would like to thank the Studienstiftung des Deutschen Volkes for supporting my work with a scholarship which allowed me to pursue my studies much more effectively and granted my a degree of flexibility otherwise impossible.

Furthermore, I am very grateful to all others who have accompanied me during my doctoral project. Because there are so many who have helped me one way or another, it is impossible to list everybody - please forgive me if I have neglected to mention you here. Thank you to the colleagues at the Bamberg Chair of Statistics and Econometrics and the Berlin Chair of Applied Statistics. Everybody who was written a dissertation knows that it is not just the work you do by yourself but also the discussions, coffee breaks, and lunch time talks which drive the research project and make it possible to continue despite many setbacks. A special thank you goes to my office mate Rachael who was forced to hear all my complaints, half-baked ideas, and was (hopefully) happy to give her input on every issue imaginable. I would also like to thank Nora, who, among many other things, gave me some direly needed breathing room during the final stretch of my dissertation.

I would like to express a special thank you to Soon-Ja Apeldorn, who helped me to understand the millions of little things that are necessary to survive in Germany and without which I would not have been able to study. I also owe a lot of gratitude to Prof. Dr. Ulrich Rendtel, who was the first to show me as an undergrad student what the university culture in Germany was like and who continued to support me every time I encountered an unexpected problem in Berlin and beyond.

Being in Germany was most likely not something that my parents envisioned for me, but in the end it was their trust in me that made it possible to send their daughter to the other end of the world. Thanks to their unwavering patience and emotional support, I was able to finish my doctorate. Thank you, mom 태은순 and dad 이원재.

Finally, thank you Benedikt for giving your input and editor's opinions on what social scientists may consider less than exciting research and writing, for incessantly asking "Ok, but what does that mean?", for your boundless emotional support, love, trust, and simply for being a part of my life.

Publication List

The publications listed below are the result of the research carried out in this thesis titled, "Recent Advances in Small Area Estimation of Economic and Poverty Indicators using Traditional and Alternative Data".

1. Lee, Y., Rojas-Perilla, N., Runge, M., and Schmid, T. (2022) **Variable selection using conditional AIC for linear mixed models with data-driven transformations**, *Statistics and Computing, 33(27), pp. 1-17,* doi: https://doi.org/10.1007/s11222-022-10198-9. Accepted and published.

2. Lee, Y. (2023) **Estimation of the consumer price index with regional weights using small area estimation methods: A case study of Germany**, *Working paper*, submitted.

3. Lee, Y., Schmid, T., and Würz, N. (2024) **Small area estimation using geospatial data based on transformed two-fold nested error regression models**, *Working paper*, to be submitted.

Contents

Introduction

Fostering effective development policy is an important concern not only for developing countries but also for developed countries. There are multiple aspects which must be considered for setting specific goals of development. In 2015, the United Nations (UN) created 17 sustainable development goals (SDGs) (United Nations, 2023). However, how important each goal is differs for each country depending on its situation. For countries suffering from extreme poverty, goals that are directly connected to poverty have a high priority such as *no poverty* (SDG 1), *zero hunger* (SDG 2), and *good health and well-being* (SDG 3). Meanwhile, developed countries aim to reduce inequality and achieve sustainable development. Therefore, these countries focus more on goals like *reduced inequalities* (SDG 10), *sustainable cities and communities* (SDG 11), and *responsible consumption and production* (SDG 12). Effective policies that are tailored to each country's particular context are essential to achieve not only the SDGs, but also further goals that drive economic and social development. Formulating effective policy starts with analyzing the current economic and social situation and tracking any changes. For their analyses, countries and organizations calculate and regularly publish important economic and poverty indicators such as gross domestic product (GDP), price indices, and head count ratio (HCR) (Foster et al., 1984).

To provide reliable indicators, countries and organizations frequently conduct several extensive surveys such as household income and expenditure surveys (International Labour Organization, 2024) or the demographic and health survey (The DHS Program, 2023). Based on collected survey data, they produce reliable indicators commonly at the national level. This information is then used to formulate policies for the entire country. However, sometimes indicators at a regionally or socio-demographically disaggregated level (e.g., districts, municipalities, or ethnic groups) are required to make a decision on where to allocate resources effectively. Furthermore, one single policy for the entire country cannot solve region-specific problems. Not considering regional differences can results in regionally unbalanced developments. To avoid this problem, it is important to understand regional differences and implement region-specific policies based on reliably estimated indicators in each region. Therefore, it is essential to have reliable indicators not only at the national or the large regional level but also at disaggregated levels. Unfortunately, because of small sample sizes and non-observed areas, it is often impossible to produce the reliable indicators of interest at a granular regional level using existing survey datasets. As these surveys are designed to produce the indicators at a national or a larger regional level, there are enough observations at these levels to estimate indicators with an acceptable level of uncertainty but not at a disaggregated level.

We can solve this problem by conducting new surveys which are designed such that there

are enough observations at the disaggregated level of interest. However, this is practically impossible due to high financial and time costs. Small area estimation (SAE) methods tackle this problem and suggest statistical solutions without further costs. Small area refers to regionally disaggregated areas or socio-demographically disaggregated groups in which the estimation of indicators with existing survey datasets is either impossible or unreliable due to small sample sizes. SAE methods use model-based approaches and combine an existing survey dataset with an auxiliary dataset to produce reliable estimates of indicators at disaggregated, i.e., small area levels (Pfeffermann, 2013; Rao and Molina, 2015; Tzavidis et al., 2018; Morales et al., 2021). Models used in SAE methods can be divided into two groups depending on which level the model is defined: (1) unit level models and (2) area level models.

SAE methods with unit level models require two datasets: a relevant survey dataset at the unit level, which is normally the household, and an additional auxiliary dataset. The survey data must include possible independent variables and the dependent variable, which is commonly an income or an expenditure variable, for the estimation of economic and poverty indicators. The auxiliary dataset must include dependent variables either at the unit level or at the area level. When a SAE approach aims to estimate nonlinear indicators, it usually requires unit level auxiliary data including the variables of all households in the population (e.g., the empirical best predictor (EBP) approach by Molina and Rao (2010)). For the estimation of linear indicators, the auxiliary data does not necessarily need to be at the unit level but the area level data should be without errors (Battese et al., 1988).

Area level SAE approaches, meanwhile, build the models at the targeted small area level (e.g., the Fay-Herriot model by Fay and Herriot (1979)). Therefore, the models require data which is aggregated at the area level. In area level models, a direct estimator of the target indicator is used as dependent variable. The dependent variable is usually obtained from a relevant survey dataset by using a direct estimation method, such as the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952). To fit the model, independent variables at area level in the auxiliary data are required and these variables are assumed to be without errors. Traditionally, census datasets are frequently used as auxiliary datasets for both unit level and area level SAE approaches. Since these datasets contain all households in the population, they satisfy the condition of unit level auxiliary data. Furthermore, the aggregation of variables from the datasets provides error free area level auxiliary information. Consequently, a census is a useful auxiliary data source not only for unit level models but also for area level models.

Since SAE methods combine two datasets, the datasets must be from the same year or at least from similar years in order for them to accurately reflect the current population. However, it is challenging to access a census dataset which represents the current population due to multiple reasons. First, due to data protection laws, it is generally difficult to access a census dataset because it includes the information of all households in the population. Second, the most recent census data may be already outdated as the census is mostly carried out only every ten years. An outdated census does not accurately represent a country's current population, especially in developing countries experiencing fast dynamic structural changes and internal conflicts. Recently, researchers have started to suggest the use of geospatial data as an alternative to census data for the estimation of economic and poverty indicators. Geospatial data

is publicly available, up to date, and covers all geographical regions. Additionally, recent empirical studies show that geospatial variables are highly correlated to the measure of household economic status, such as income or expenditure (Steele et al., 2017; Yeh et al., 2020; Newhouse, 2023). Evidently, they represent an interesting data source which can replace the census data in the SAE context.

This thesis investigates how SAE results can be improved with traditional and alternative data. To obtain efficient and reliable SAE results, it is important not only to find appropriate datasets but to also have model assumptions satisfied. Adequately transforming the dependent variable of the model is one of the most common solutions for unfulfilled model assumptions (Box and Cox, 1964; Yang, 2006). An adequate transformation also enables a more precise estimation of specific types of indicators (e.g., ratio type indicators) with area level models. For these reasons, transformations play an important role in SAE research (Slud and Maiti, 2006; Sugasawa and Kubokawa, 2015, 2017; Rojas-Perilla et al., 2020). Still, some problems remain. One example would be the variable selection for the models with data-driven transformations and the correction of the bias caused by the back-transformation. In addition to the main focus on the use of traditional and alternative data in SAE, transformation problems also play a central role in this thesis.

The first two chapters focus on the estimation of indicators with traditional datasets. Chapter 1 focuses on developing countries. For developing countries, poverty continues to be a major problem. Therefore, it is important to reliably estimate poverty indicators. In Chapter 1, poverty indicators in Mexico are estimated with the national survey of household income and expenditure in 2010 and the census from the same year as auxiliary data based on the EBP approach which is one of the most popular unit level SAE approaches. To achieve precise estimates, finding an optimal set of variables and having the model assumptions fulfilled is a central point. In practice, however, (a) the optimal model is unknown and (b) the distributional model assumptions are not fulfilled. Although these problems commonly appear together, researchers tend to treat them individually by (a) finding an optimal model based on the (conditional) Akaike information criterion (AIC) (Akaike, 1973; Vaida and Blanchard, 2005) and (b) applying transformations on the dependent variable. However, the optimal model depends on the transformation and vice versa. Chapter 1 solves these problems simultaneously by using Jacobian adjusted conditional AIC so that the estimated indicators are more accurate and reliable.

Unlike developing countries, extreme poverty is no longer an issue for most developed countries. These countries concentrate on inequality problems and sustainable development. Chapter 2 focuses on the estimation of the regional consumer price index (CPI) with regional product weights in Germany using traditional data (e.g., income and consumption survey (RDC, 2018a)). Regional CPI illustrates the regional differences in price level of consumed goods by households and can therefore serve as a indicator for regional inequality. Countries publishing the regional CPI, including Germany, use consistent national product weights (Statistisches Bundesamt, 2021b; BLS, 2023b; Statistics Canada, 2023) which show how much money is comparatively spent nationally on each product. However, since regional spending on goods differs from national spending, the use of national product weights is inaccurate to

estimate regional CPI (Kosfeld et al., 2009). Because estimating reliable regional spending is difficult due to the small sample sizes, Chapter 2 suggests the construction of regional product weights by estimating the regional household expenditure based on multivariate Fay-Herriot (MFH) models (Benavent and Morales, 2016). The MFH models allow for a more flexible covariance structure so that the correlation between expenditure on different products can be captured in the covariance matrix. In other words, using MFH models allows for capturing the reality that spending money on a product influences how much money is spent on other products - if people buy bread they also tend to buy butter. Through this more realistic model choice, the regional product weights are estimated more accurately.

Whereas Chapter 1 and 2 deal with SAE using traditional data, Chapter 3 focuses on using alternative auxiliary data. Chapter 3 investigates how to use the geospatial data as alternative data source for the estimation of the poverty indicators in Mozambique, one of the poorest countries in the world based on GDP per capita (IMF, 2023). Unlike the traditional data collected at the household level, geospatial data is extracted from satellite imagery divided into multiple cells called grids. Therefore, geospatial data is recorded at the grid level. A grid is geographically larger than a household and smaller than a common small area such as a district. Since the majority of SAE researchers have, until now, used either a unit level model or a area level model, there is no clear scholarly consensus on which model can optimize the advantages of geospatial data (Masaki et al., 2022; Corral et al., 2021). An appropriate model for the geospatial data must take into account this hierarchical data structure (Erciulescu et al., 2019). Chapter 3 introduces the two-fold nested error regression model (Torabi and Rao, 2014) with transformations for geospatial data. By using the two-fold model, the information loss is reduced and the data structure is captured in the random effects. Moreover, applying transformations allows for a correction of the violation in distributional assumptions and the estimation of ratio type indicators, such as the HCR. Chapter 3's transformed two-fold model shows efficiency gains compared to the standard transformed Fay-Herriot model.

Chapter 1

Variable selection using conditional AIC for linear mixed models with data-driven transformations

Abstract

When data analysts use linear mixed models, they usually encounter two practical problems: a) the true model is unknown and b) the Gaussian assumptions of the errors do not hold. While these problems commonly appear together, researchers tend to treat them individually by a) finding an optimal model based on the conditional Akaike information criterion (*cAIC*) and b) applying transformations on the dependent variable. However, the optimal model depends on the transformation and vice versa. In this paper, we aim to solve both problems simultaneously. In particular, we propose an adjusted *cAIC* by using the Jacobian of the particular transformation such that various model candidates with differently transformed data can be compared. From a computational perspective, we propose a stepwise selection approach based on the introduced adjusted *cAIC*. Model-based simulations are used to compare the proposed selection approach to alternative approaches. Finally, the introduced approach is applied to Mexican data to estimate poverty and inequality indicators for 81 municipalities.

Keywords: Box-Cox transformation, Empirical best predictor, Indicators, Small area estimation

1.1 Introduction

The linear mixed model is a broadly used statistical model for analyzing clustered or longitudinal data. When data analysts use these models, they often face two practical problems: a) the true model for explaining the response variable is unknown and b) the model assumptions, especially the Gaussian assumptions of the error terms, are violated.

As the true model is unknown, data analysts find suitable/optimal models for explaining the dependent variable by using variable selection procedures. One popular approach in this context is the Akaike information criterion (AIC) introduced by Akaike (1973). For linear mixed models, there are different versions of AIC (Müller et al., 2013). They can be divided into two groups: marginal types of AIC ($mAIC$) and conditional types of AIC ($cAIC$). The $mAIC$ is the common AIC for linear mixed models which uses marginal density and is one of the most widely used selection criteria (Müller et al., 2013). However, the $mAIC$ is only

appropriate when the model parameters are fixed (Burnham and Anderson, 2002) and the use of $mAIC$ as selection criterion is problematic for linear mixed models (Han, 2013). Vaida and Blanchard (2005) introduced the $cAIC$ as a more proper selection criterion for linear mixed models. $cAIC$ uses the conditional density in contrast to $mAIC$. Vaida and Blanchard (2005) derive $cAIC$ in case that the (scaled) covariance matrix of random effects is known and recommend to use a plug-in estimator for the covariance matrix of the random effects in practice. Liang et al. (2008) derive a more general $cAIC$ that accounts for the estimation of the covariance matrix of the random effects. However, their conditional AIC can be computationally demanding in situations with large sample sizes and many potential variables (Greven and Kneib, 2010).

Linear mixed models regularly rely on parametric assumptions such as normality for the random effects and the error terms. These assumptions may be violated in many applications, for instance, with skewed variables like consumption or income. One possible way to tackle this issue is to use robust mixed models. Such models are robust in various aspects, including the violation of the Gaussian assumptions. They allow more flexible distributions (Verbeke and Lesaffre, 1997; Zhang and Davidian, 2001; Sinha and Rao, 2009) or apply a Bayesian framework (Rosa et al., 2003; Lachos et al., 2009). Jiang (2019) gives an overview of further models which deal with this problem. Another way to solve this problem is to apply fixed logarithmic or data-driven transformations for the dependent variable. The latter transformations are generally based on an adaptive transformation parameter λ that depends on the particular shape of the data. Among different data-driven transformations, the Box-Cox transformation (Box and Cox, 1964) is widely used, as it includes various power transformations and the logarithmic transformation as a special case. Gurka et al. (2006) extend the use of the Box-Cox transformation to linear mixed models. They apply the residual maximum likelihood (REML) approach to estimate the transformation parameter λ from the data based on linear mixed model with fixed auxiliary variables.

However, the optimal data-driven transformation depends on the fixed model and the optimal model depends on the selected data-driven transformation. In particular, to select the optimal data-driven transformation parameter λ by the REML approach, the linear mixed model should be fixed; and to perform a variable selection based on the $cAIC$, the dependent variable should be fixed using an appropriate (data-driven) transformation parameter λ. A first naive approach which is typically used in applications would be to perform the transformation and variable selection in a specific order. First, find an appropriate working model on the original/untransformed scale and keep this fixed when selecting the optimal data-driven transformation parameter. However, this may not offer the best way to the variable selection as the selected variables are not optimal on the transformed scale. In this paper, we aim to find the optimal model and the optimal transformation parameter simultaneously. This would allow for enjoying the advantages of both data-driven transformations and the optimal model for the transformed data.

Hoeting and Ibrahim (1998) and Hoeting et al. (2002) discuss methods for transformation and variable selection based on posterior probabilities in linear models. They focus on change-point transformations to transform the predictors of the linear model. Bunke et al.

(1999) discuss the selection of the optimal transformation and the optimal model based on cross validation for the nonlinear model. To the best of our knowledge, none of the existing literature provides a joint solution when variable selection based on the $cAIC$ and estimation of the data-driven transformation parameter are simultaneously applied to linear mixed models. From a theoretical perspective, we present an approach to concurrently choose the optimal linear model and the optimal transformation parameter. Since the $cAIC$ is scale dependent, we can not directly compare different models with differently transformed response variables. Therefore, we adjust the $cAIC$ using the Jacobian of the corresponding data-driven transformation such that different model candidates with differently transformed response variables can be compared. Although the paper focuses on the Box-Cox transformation as a particular data-driven transformation, the proposed approach is applicable to data-driven transformations in general. From a computational perspective, we provide a step-wise selection approach based on the proposed adjusted $cAIC$.

The structure of the paper is as follows: In Section 1.2 we provide an overview of linear mixed models and the $cAIC$. In Section 1.3, we derive the Jacobian adjusted $cAIC$ for transformed linear mixed models and introduce the step-wise selection approach. In Section 1.4, we examine the performance of the proposed selection approach by using model-based simulations. In Section 1.5, the proposed selection approach is applied to data from Guerrero in Mexico for estimating poverty and inequality indicators at municipal level. Finally, we discuss our results and further directions of research in Section 1.6.

1.2 Variable selection using conditional AIC for linear mixed models

In this section we briefly introduce the existing variable selection methods for linear mixed models. In Section 1.2.1, we present a general notation of linear mixed models and in Section 1.2.2, we introduce and compare the $cAIC$ by Vaida and Blanchard (2005) and Liang et al. (2008).

1.2.1 The linear mixed model

Assume there is a finite population divided into D clusters. Let y_i be a vector of the response variable for the i-th cluster for $i = 1, \cdots, D$, which is modeled with a linear mixed model

$$y_i = X_i\beta + Z_i u_i + \varepsilon_i.$$

N_i is the cluster size of the i-th cluster, X_i and Z_i are known $N_i \times p$ and $N_i \times q$ design matrices for the fixed and random effects, β includes p fixed effects, u_i is a vector of q random effects, and ε_i is a vector of errors in the i-th cluster. u_i and ε_i are assumed to be independent and normally distributed

$$u_i \sim \mathcal{N}(0, G),\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{N_i}),$$

with I_{N_i}, the $N_i \times N_i$ identity matrix. G is the $q \times q$ covariance matrix of random effects in the i-th cluster and depends on a set of variance components η. Let $N = \sum_{i=1}^{D} N_i$ be the population size and $\theta = (\beta, \sigma, \eta)$ be the vector of parameters in the model. The model is described for the population as follows

$$y = X\beta + Zu + \varepsilon, \tag{1.1}$$

where $X = (X_1^T, \cdots, X_D^T)^T$ is a $N \times p$ matrix, $Z = \text{diag}(Z_1, \cdots, Z_D)$ is $N \times r$ block-diagonal matrix with $r = D \cdot q$, $u = (u_1^T, \cdots, u_D^T)^T$ and $\varepsilon = (\varepsilon_1^T, \cdots, \varepsilon_D^T)^T$. ε and u are independent and normally distributed with $E(\varepsilon) = E(u) = 0$, $Var(\varepsilon) = \sigma^2 I_N$ and $Var(u) = G_0$, where $G_0 = diag_D(G)$ is block-diagonal matrix with D blocks of G on the diagonal. As $u \sim \mathcal{N}(0, G_0)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_N)$, the covariance matrix of y is given by

$$Cov(y) = V = \sigma^2 I_N + ZG_0 Z^T.$$

1.2.2 Conditional Akaike information criterion for linear mixed models

Assume there are P possible explanatory variables in the data. Since the number of all possible combinations of P variables is $M = 2^P$, there are M possible model candidates which can be fitted to the data. In order to find the optimal model among them, the variable selection should be performed based on an appropriate selection criterion. In this study, we focus on the variable selection based on the $cAIC$ for linear mixed models. While this study focuses on the $cAIC$, the $mAIC$ is briefly explained first to provide a better understanding of the $cAIC$.

The $mAIC$ is derived from the Kullback-Leibler (K-L) divergence between the density of the true model and the density of a candidate model (Akaike, 1973). Assume that the true model has the same form as Equation (1.1) with true parameters. The vector of true parameters is denoted by $\theta_0 = (\beta_0, \sigma_0, \eta_0)$. Let $f(\cdot)$ be the density function of the true generating model and $g(\cdot|\theta)$ be the density of the approximating model with model parameters θ for fitting the data. If the true distribution f belongs to the class of model candidates and $\theta = \theta_0$ then $g(\cdot|\theta_0) = f(\cdot)$. The $mAIC$ measures the K-L divergence between $f(\cdot)$ and $g(\cdot|\theta)$.

The idea behind the $cAIC$ derivation is the same as for the $mAIC$. While the $mAIC$ measures the K-L divergence between two marginal densities, $cAIC$ measures the K-L divergence between the true conditional density and the conditional density of a model candidate. The true conditional density is denoted by $f(\cdot|u_0)$ with the true random effects (u_0) and the conditional density of a model candidate is denoted by $g(\cdot|\theta, u)$. Let y^* be generated from the true conditional density and y be the observed data, also from the true conditional density. They are independent conditional on random effects, which means that y^* and y share the random effects and only differ in error terms (i.e., $y^* = X\beta + Zu + \varepsilon^*$ and $y = X\beta + Zu + \varepsilon$ with $\varepsilon^* \sim N(0, \sigma^2 I_N)$ and $\varepsilon \sim N(0, \sigma^2 I_N)$). The K-L divergence between $f(y^*|u_0)$ and $g(y^*|\theta, u)$ with respect to $f(y^*|u_0)$ is defined by

$$\begin{aligned} I[(\theta_0, u_0), (\theta, u)] =& E_{f(y^*|u_0)}\left[\log \frac{f(y^*|u_0)}{g(y^*|\theta, u)}\right] \\ =& E_{f(y^*|u_0)}[\log f(y^*|u_0)] - E_{f(y^*|u_0)}[\log g(y^*|\theta, u)]. \end{aligned}$$

The discrepancy between the conditional generating model and the conditional approximation model is given by

$$d[(\theta_0, u_0), (\theta, u)] = E_{f(y^*|u_0)}[-2\log g(y^*|\theta, u)].$$

By using the given definition of the discrepancy, the K-L divergence can be written as follows

$$2I[(\theta_0, u_0), (\theta, u)] = 2E_{f(y^*|u_0)}[\log f(y^*|u_0)] + d[(\theta_0, u_0), (\theta, u)].$$

Since $2E_{f(y^*|u_0)}[\log f(y^*|u_0)]$ does not depend on θ and u from the approximating model, the ranking of candidate models based on $d[(\theta_0, u_0), (\theta, u)]$ is equivalent to the ranking of candidates based on $2I[(\theta_0, u_0), (\theta, u)]$. Therefore, the fitted candidate models can be evaluated by using the discrepancy with $\hat{\theta}$ and $\hat{u}$,

$$d[(\theta_0, u_0), (\theta, u)] = d[(\theta_0, u_0), (\theta, u)]|_{\theta=\hat{\theta}, u=\hat{u}},$$

where $\hat{\theta}$ includes the estimates of model parameters (i.e. $\hat{\theta} = (\hat{\beta},\ \hat{\sigma},\ \hat{\eta})$) and $\hat{u} = E(u|\hat{\theta},\ y)$ contains the predicted random effects based on the empirical Bayes estimation. Hence, the selection problem based on K-L divergence can be solved by comparing $d[(\theta_0, u_0), (\theta, u)]|_{\theta=\hat{\theta}, u=\hat{u}}$ values of the candidate models. As the model parameters and random effects are estimated based on observed data, the expected estimated discrepancy should be used as the selection criterion (Burnham and Anderson, 2002). This is also often denoted as conditional Akaike Information (cAI) (Vaida and Blanchard, 2005; Liang et al., 2008; Han, 2013)

$$cAI = E_{f(y,u)}E_{f(y^*|u)}[-2\log g(y^*|\hat{\theta}, \hat{u})].$$

$-\log g(y|\hat{\theta}, \hat{u})$ is a biased estimator of $E_{f(y,u)}E_{f(y^*|u)}[-\log g(y^*|\hat{\theta}, \hat{u})]$. As a consequence, the $cAIC$ consists of the conditional log-likelihood and the bias correction term K

$$cAIC = -2\log g(y|\hat{\theta}, \hat{u}) + 2K,$$

where

$$\log g(y|\hat{\theta}, \hat{u}) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - \hat{y})^T(y - \hat{y}),$$

and $\hat{y}$ is the fitted vector $\hat{y} = X\hat{\beta} + Z\hat{u}$.

Vaida and Blanchard (2005) derive two different bias correction terms under different assumptions. When σ^2 and G_0 are assumed to be known, the K equals ρ, which is the effective degrees of freedom (Hodges and Sargent, 2001)

$$K_a = \rho = tr\left[\begin{pmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ + \sigma^{-2}G_0 \end{pmatrix}^{-1}\begin{pmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ \end{pmatrix}\right].$$

When it is assumed that σ^2 is unknown and $\sigma^{-2}G_0$ is known, K is calculated by

$$K_{MLE} = \frac{N(N-p-1)}{(N-p)(N-p-2)}(\rho+1) + \frac{N(p+1)}{(N-p)(N-p-2)}. \tag{1.2}$$

The detailed derivation of K_a and K_{MLE} can be found in Vaida and Blanchard (2005).

Vaida and Blanchard (2005) derive the $cAIC$ under the assumption that G_0, the covariance matrix of the random effects, or $\sigma^{-2}G_0$, the scaled covariance matrix of the random effects, are known. However, in practice they are usually unknown. In the case of the unknown random effects covariance matrix, Vaida and Blanchard (2005) suggest to use K_{MLE} for the $cAIC$ with the estimated $\sigma^{-2}G_0$, since the derivation of the bias correction term for the case of unknown $\sigma^{-2}G_0$ is analytically complicated and the effect of estimation can be asymptotically ignored.

Liang et al. (2008) propose a general $cAIC$ for known σ^2, regardless of whether the covariance of random effects are known or unknown. Under these assumptions, Liang et al. (2008) derive the bias correction term using the first derivatives of $\hat{y}$ subject to y. In their technical report, they also derive an additional bias correction term for $cAIC$ assuming more realistically that neither σ^2 nor the covariance of random effects are known. In practice, the true value of σ^2 and the true G_0 are usually unknown. Therefore, it seems reasonable to use the $cAIC$ of Liang et al. (2008). However, Liang et al. (2008) show in the simulation part that their bias correction term is close to K_a and it is also shown in their technical report that the bias correction term under more realistic assumptions is close to K_{MLE}. Moreover, Greven and Kneib (2010) point out that the use of $cAIC$ by Liang et al. (2008) as a selection criterion poses severe computational difficulties, since the calculation of the bias correction term of Liang et al. (2008) requires at least N additional model fits to calculate derivatives. If there are M different model candidates, at least $N \times M$ model fits are required to calculate $cAIC$ derived by Liang et al. (2008), which is hard to implement for large N and M. As a result, this study focuses on the $cAIC$ of Vaida and Blanchard (2005), and in particular on the $cAIC$ with K_{MLE} that allows for unknown σ^2. The optimal model is the model which has the minimum value of $cAIC$ among all M model candidates.

1.3 Variable selection for linear mixed models with transformations

In this section, we propose a step-wise variable selection approach for linear mixed models which allows comparing model candidates with differently transformed response variables. First, we give a general notation of linear mixed models with the Box-Cox transformation. Although the paper focuses on the Box-Cox transformation as a particular transformation, the proposed approach is applicable to data-driven transformations in general. In Section 1.3.2, we derive the Jacobian adjusted $cAIC$ based on $cAIC$ by Vaida and Blanchard (2005), which can compare model candidates with differently transformed data. In Section 1.3.3, we introduce a bootstrap method to estimate the bias correction term for Jacobian adjusted $cAIC$. From a computational perspective, we suggest to use step-wise selection with adjusted $cAIC$ in Section 1.3.4.

1.3.1 Linear mixed models with transformations

Assume that the original y variable is non-normal and there exists a transformation parameter of the Box-Cox transformation for which the transformed data follows the Gaussian assumption.

The one-to-one Box-Cox transformation (Box and Cox, 1964) of y is defined by

$$T_\lambda(y_{ij}) = \begin{cases} \frac{(y_{ij}+s)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(y_{ij}+s) & \text{if } \lambda = 0, \end{cases} i = 1, ..., D \text{ and } j = 1, ..., N_i, \tag{1.3}$$

where λ denotes the transformation parameter which has to be estimated and s denotes the shift parameter $s = |\min(y)| + 1$ only when $\min(y) < 0$. Let $\widetilde{y}$ be the vector of transformed y. Then, $\widetilde{y}$ is modeled as

$$T_\lambda(y) = \widetilde{y} = X\beta + Zu + \varepsilon \tag{1.4}$$

with $u \sim \mathcal{N}(0, G_0)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_N)$. The covariance matrix of the transformed y is

$$Cov(\widetilde{y}) = V = \sigma^2 I_N + ZG_0Z^T.$$

Gurka et al. (2006) use the REML approach to estimate λ, as the REML approach is recommended when the focus is the estimation of variance components (Verbeke and Molenberghs, 2000). Moreover, Rojas-Perilla et al. (2020) compare the REML estimator of λ with other estimators and show that the REML approach has a smaller variability than alternative estimators. Accordingly, the optimal λ is estimated in this study with the REML approach. The optimal λ maximizes the residual log-likelihood function of a given model. However, the estimated optimal λ is only optimal for the given model. This means that each model candidate has its own optimal λ. As we do not know which model candidate is the optimal and which λ is the optimal for the corresponding model, we should select the model and the λ concurrently.

To simultaneously select the best model based on $cAIC$ and obtain the optimal λ, we estimate it for each potential model in a first step. With P possible x variables there are $M = 2^P$ model candidates. The m-th model is defined by

$$\begin{aligned} T_{\lambda_m}(y) = \widetilde{y}^{\,(m)} &= X^{(m)}\beta + Zu + \varepsilon, \\ m &= 1, ..., M, \end{aligned} \tag{1.5}$$

where $X^{(m)}$ is the design matrix of the m-th model and λ_m is the optimal transformation parameter for the m-th model. Based on the model in Equation (1.5) the optimal transformation parameter is estimated using the REML approach and $\hat{\lambda}_m$ denotes the estimated optimal transformation parameter for the m-th model. Further details about the estimation of λ_m using the REML approach are explained in Gurka et al. (2006). In the second step, all model candidates with their own $\hat{\lambda}_m$ should be compared. However, AIC-type criteria cannot compare models with differently transformed target variable (Burnham and Anderson, 2002). Therefore, an adjustment with the Jacobian to the $cAIC$ should be performed first such that these M different models can be compared.

1.3.2 Jacobian adjusted *cAIC* for linear mixed models

Assume that $f(\cdot|u_0)$ is the true conditional density function with the true model parameters θ_0 and the true random effects u_0, while $g(\cdot|\theta, u)$ denotes the conditional density of an approximating model. Let $\widetilde{y}^* = X\beta + Zu + \varepsilon^*$ be a realization from the true conditional density function with $\varepsilon^* \sim N(0, \sigma^2)$. Then, the cAI for the transformed model is given by

$$cAI = E_{f(\widetilde{y},u)} E_{f(\widetilde{y}^*|u)}[-2 \log g(\widetilde{y}^*|\hat{\theta}, \hat{u})],$$

where $\hat{\theta}$ is the vector of estimated model parameters and $\hat{u}$ is the vector of predicted random effects. $-\log g(\widetilde{y}|\hat{\theta}, \hat{u})$ is a biased estimator of $E_{f(\widetilde{y},u)} E_{f(\widetilde{y}^*|u)}[-\log g(\widetilde{y}^*|\hat{\theta}, \hat{u})] = 0.5 \cdot cAI$. The bias is obtained by

$$bias = E_{f(\widetilde{y},u)}[-\log g(\widetilde{y}|\hat{\theta}, \hat{u})] - 0.5 \cdot cAI.$$

To obtain an unbiased estimator of $0.5 \cdot cAI$, the bias correction term (BC) should be added as follows

$$\begin{aligned} BC =& - E_{f(\widetilde{y},u)}[-\log g(\widetilde{y}|\hat{\theta}, \hat{u})] + 0.5 \cdot cAI \\ =& E_{f(\widetilde{y},u)}[\log g(\widetilde{y}|\hat{\theta}, \hat{u})] - E_{f(\widetilde{y},u)} E_{f(\widetilde{y}^*|u)}[\log g(\widetilde{y}^*|\hat{\theta}, \hat{u})] \\ =& E\left[\frac{1}{2\sigma^2}[(\widetilde{y}^* - \widehat{\widetilde{y}})^T(\widetilde{y}^* - \widehat{\widetilde{y}}) - (\widetilde{y} - \widehat{\widetilde{y}})^T(\widetilde{y} - \widehat{\widetilde{y}})]\right], \end{aligned} \tag{1.6}$$

where $\widehat{\widetilde{y}} = X\hat{\beta} + Z\hat{u}$.

Under the assumption that σ^2 is unknown, the BC in Equation (1.6) can be replaced by K_{MLE} from Equation (1.2). Consequently, the $cAIC$ for the transformed model is given by

$$cAIC = -2 \log g(\widetilde{y}|\hat{\theta}, \hat{u}) + 2K_{MLE}, \tag{1.7}$$

where

$$\log g(\widetilde{y}|\hat{\theta}, \hat{u}) = -\frac{N}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\widetilde{y} - \widehat{\widetilde{y}})^T(\widetilde{y} - \widehat{\widetilde{y}}).$$

However, this $cAIC$ of the transformed model cannot be used to compare differently transformed model candidates. The $cAIC$ measures the K-L distance between the true conditional density and a conditional density of a model candidate. In the case of linear mixed models without a transformation, the optimal model can be chosen using the $cAIC$ by Vaida and Blanchard (2005), since all model candidates have the same response variable y. The model with the smallest distance (i.e., the smallest $cAIC$) is the optimal model among all candidates. However, for linear mixed models with a transformation we estimate for each model candidate its own optimal transformation parameter. As the transformation parameter differs from model to model, the transformed y differs too. Consequently, the response variables of the model candidates are no longer the same (i.e., $\widetilde{y}^{(1)} \neq \widetilde{y}^{(2)} \neq \cdots \neq \widetilde{y}^{(M)}$). Therefore, the $cAIC$ in Equation (1.7) of a model candidate is in fact not the distance of the model from the true density of y, but the distance from the true density of $\widetilde{y}$. As $\widetilde{y}$ differs from candidate to candidate and $cAIC$ is scale dependent, the model candidates cannot be compared with the $cAIC$. To allow for comparing model candidates using the $cAIC$, it needs to be adjusted, so that the adjusted $cAIC$ of a model candidate measures the divergence of the model from the true density of y. Akaike (1978) shows that this adjustment can be done by adding the Jacobian of the transformation to the AIC value of time

series models.

The Jacobian adjusted $cAIC$ denoted by $JcAIC$ is derived from the K-L divergence between the true conditional density and the model conditional density of the original y, and not of the transformed y. To define the K-L divergence between the true and a candidate model of y, the true and model conditional densities of y should be defined. As we know the conditional densities of the transformed y, the conditional densities of y can be derived by multiplying the Jacobian of the transformation. Let $h(y|u_0)$ be the true conditional density of y and $l(y|\theta, u)$ the conditional model density, which are defined with the Jacobian of the Box-Cox transformation $J(\lambda, y)$ as

$$\begin{aligned} h(y|u_0) =& f(\widetilde{y}|u_0) \cdot J(\lambda, y), \\ l(y|\theta, u) =& g(\widetilde{y}|\theta, u) \cdot J(\lambda, y), \end{aligned} \tag{1.8}$$

where

$$J(\lambda, y) = \frac{\partial \widetilde{y}}{\partial y} = \prod_{i=1}^{D} \prod_{j=1}^{N_i} \frac{\partial \widetilde{y}_{ij}}{\partial y_{ij}} = \prod_{i=1}^{D} \prod_{j=1}^{N_i} (y_{ij} + s)^{\lambda - 1}. \tag{1.9}$$

Let y^* be a realization of the true conditional density $h(y|u)$ and $\widetilde{y}^*$ be the vector of transformed y^*. Then, the K-L divergence between conditional densities of y^* becomes

$$\begin{aligned} I[(\theta_0, u_0), (\theta, u)] =& E_{h(y^*|u)} \left[\log \frac{h(y^*|\theta_0, u_0)}{l(y^*|\theta, u)} \right] \\ =& E_{h(y^*|u)}[\log h(y^*|\theta_0, u_0)] - E_{h(y^*|u)}[\log l(y^*|\theta, u)]. \end{aligned}$$

The discrepancy is defined by $d[(\theta_0, u_0), (\theta, u)] = E_{h(y^*|u)}[-2 \log l(y^*|\theta, u)]$. Therefore, the K-L divergence can be formulated using discrepancies as follows

$$2I[(\theta_0, u_0), (\theta, u)] = 2E_{h(y^*|u)}[\log h(y^*|\theta_0, u_0)] + d[(\theta_0, u_0), (\theta, u)].$$

The ranking of $d[(\theta_0, u_0), (\theta, u)]$ is equivalent to the ranking of $2I[(\theta_0, u_0), (\theta, u)]$, since the first term $2E_{h(y^*|u)}[\log h(y^*|\theta_0, u_0)]$ is constant for all model candidates. The Jacobian adjusted cAI ($JcAI$) is

$$JcAI = E_{h(y,u)} E_{h(y^*|u)}[-2 \log l(y^*|\hat{\theta}, \hat{u})].$$

$-\log(l(y|\hat{\theta}, \hat{u}))$ is a biased estimator of $0.5 \cdot JcAI$. To obtain an unbiased estimator of $0.5 \cdot JcAI$, the bias should be corrected by the following bias correction term (BC):

$$\begin{aligned} BC =& -\Big(E_{h(y,u)}[-\log(l(y|\hat{\theta}, \hat{u}))] - 0.5 \cdot JcAI \Big) \\ =& E_{h(y,u)}[\log(l(y|\hat{\theta}, \hat{u}))] - E_{h(y,u)} E_{h(y^*|u)}[\log(l(y^*|\hat{\theta}, \hat{u}))]. \end{aligned}$$

$l(y|\hat{\theta}, \hat{u})$ is defined as in Equation (1.8). Then, $l(y^*|\hat{\theta}, \hat{u})$ can be defined by $g(\widetilde{y}^*|\hat{\theta}, \hat{u}) \cdot J(\hat{\lambda}, y^*)$ using the same relation as in Equation (1.8). By inserting these terms into the BC, we get

$$\begin{aligned} BC =& E_{h(y,u)}[\log(g(\widetilde{y}|\hat{\theta}, \hat{u}) \cdot J(\hat{\lambda}, y))] - E_{h(y,u)} E_{h(y^*|u)}[\log(g(\widetilde{y}^*|\hat{\theta}, \hat{u}) \cdot J(\hat{\lambda}, y^*))] \\ =& E\Big[-\frac{N}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} (\widetilde{y} - \widehat{\widetilde{y}})^T (\widetilde{y} - \widehat{\widetilde{y}}) + \log(J(\hat{\lambda}, y)) \Big] \\ & - E\Big[-\frac{N}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} (\widetilde{y}^* - \widehat{\widetilde{y}})^T (\widetilde{y}^* - \widehat{\widetilde{y}}) + \log(J(\hat{\lambda}, y^*)) \Big]. \end{aligned} \tag{1.10}$$

The Jacobian term of y is defined in Equation (1.9) and the Jacobian term for y^* is given by $\prod_{i=1}^{D} \prod_{j=1}^{N_i} (y_{ij}^* +$

$s)^{\lambda-1}$ leading to

$$BC = E\Big[-\frac{N}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\widetilde{y}-\widehat{\widetilde{y}})^T(\widetilde{y}-\widehat{\widetilde{y}}) + (\hat{\lambda}-1)\sum_{i=1}^{D}\sum_{j=1}^{N_i}\log(y_{ij}+s))\Big]$$
$$- E\Big[-\frac{N}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\widetilde{y}^*-\widehat{\widetilde{y}})^T(\widetilde{y}^*-\widehat{\widetilde{y}}) + (\hat{\lambda}-1)\sum_{i=1}^{D}\sum_{j=1}^{N_i}\log(y_{ij}^*+s)\Big]. \quad (1.11)$$

1.3.3 Estimation of the bias correction

We propose a parametric bootstrap - following the ideas of Donohue et al. (2011) and Rojas-Perilla et al. (2020) - to estimate the BC for the $JcAIC$. The bootstrap captures not only the uncertainty due to the estimation of the model parameters but also the additional uncertainty due to the estimation of the transformation parameter λ (Rojas-Perilla et al., 2020). In addition, we use a resampling approach because the bootstrap variants of AIC are comparable with analytic approximations of the AIC (Donohue et al., 2011) and perform better than analytic approximations in terms of the model choice (Shang and Cavanaugh, 2008; Marhuenda et al., 2014).

The BC in Equation (1.11) consists of two expectation terms. Each expectation term is estimated by averaging the values over the B bootstrap replicates. The steps of the proposed bootstrap are as follows:

1. Estimate the optimal λ defined as $\hat{\lambda}$ using REML for the model candidate and transform the y to the $\widetilde{y}$ with the estimated $\hat{\lambda}$.
2. Fit the model in Equation (1.4) to obtain estimates of model parameters $\hat{\theta}$.
3. Generate $u^{(b)} \sim \mathcal{N}(0, \hat{G}_0)$ and $\varepsilon^{(b)} \sim \mathcal{N}(0, \hat{\sigma}^2)$ and create a bootstrap $\widetilde{y}$ using $\widetilde{y}^{(b)} = X\hat{\beta} + Zu^{(b)} + \varepsilon^{(b)}$.
4. Refit the model with the bootstrap sample $\widetilde{y}^{(b)}$ and obtain the bootstrap estimates of the model parameters $\hat{\theta}^{(b)}$ and $\hat{u}^{(b)}$.
5. Calculate the second expectation term of the BC for each bootstrap using $\hat{\theta}^{(b)}$ and $\hat{u}^{(b)}$. The unobserved (true) $\widetilde{y}^*$ and y^* are replaced by $\widetilde{y}$ and y respectively. Note that $\widetilde{y}$ and y are treated as realizations from the true transformed/untransformed density with corresponding $\hat{\lambda}$.
6. Back-transform $\widetilde{y}^{(b)}$ using $\hat{\lambda}$ to obtain $y^{(b)}$ on the original scale. Re-estimate $\hat{\lambda}^{(b)}$ based on $y^{(b)}$ and re-transform the $y^{(b)}$ using $\hat{\lambda}^{(b)}$. The re-transformed bootstrap $y^{(b)}$ is denoted by $\widetilde{y}^{(\hat{\lambda}^{(b)},(b))}$.
7. Refit the model with the bootstrap sample $\widetilde{y}^{(\hat{\lambda}^{(b)},(b))}$ and obtain the bootstrap estimates of the model parameters $\hat{\theta}^{(\hat{\lambda}^{(b)},(b))}$ and $\hat{u}^{(\hat{\lambda}^{(b)},(b))}$. Note that the estimates depend on the re-estimated transformation parameter indicated by the superscript $\hat{\lambda}^{(b)}$.
8. Calculate the first expectation term of the BC for each bootstrap using $\hat{\theta}^{(\hat{\lambda}^{(b)},(b))}$, $\hat{u}^{(\hat{\lambda}^{(b)},(b))}$, $\hat{\lambda}^{(b)}$, $\widetilde{y}^{(\hat{\lambda}^{(b)},(b))}$ and $y^{(b)}$.

The bootstrap estimate of the BC is then obtained by

$$BC = \frac{1}{B}\sum_{b=1}^{B}\Bigg[-\frac{N}{2}log\big(2\pi\hat{\sigma}^{2(\hat{\lambda}^{(b)},(b))}\big) - \frac{1}{2\hat{\sigma}^{2(\hat{\lambda}^{(b)},(b))}}\cdot\Big(\widetilde{y}^{(\hat{\lambda}^{(b)},(b))} - X\hat{\beta}^{(\hat{\lambda}^{(b)},(b))} - Z\hat{u}^{(\hat{\lambda}^{(b)},(b))}\Big)^T$$
$$\Big(\widetilde{y}^{(\hat{\lambda}^{(b)},(b))} - X\hat{\beta}^{(\hat{\lambda}^{(b)},(b))} - Z\hat{u}^{(\hat{\lambda}^{(b)},(b))}\Big) + \big(\hat{\lambda}^{(b)}-1\big)\sum_{i=1}^{D}\sum_{j=1}^{N_i}\log\big(y_{ij}^{(b)}+s\big)\Bigg]$$
$$-\frac{1}{B}\sum_{b=1}^{B}\Bigg[-\frac{N}{2}log\big(2\pi\hat{\sigma}^{2(b)}\big) - \frac{1}{2\hat{\sigma}^{2(b)}}\cdot\Big(\widetilde{y} - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\Big)^T\Big(\widetilde{y} - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\Big)$$

$$+ \left(\hat{\lambda} - 1\right) \sum_{i=1}^{D} \sum_{j=1}^{N_i} \log\left(y_{ij} + s\right) \Bigg]. \tag{1.12}$$

Then, the $JcAIC$ is estimated by

$$\begin{aligned} JcAIC &= -2\log(l(y|\hat{\theta}, \hat{u})) + 2BC \\ &= -2\log(g(\widetilde{y}|\hat{\theta}, \hat{u})) - 2\log(J(\hat{\lambda}, y)) + 2BC \end{aligned} \tag{1.13}$$

with BC defined in Equation (1.12).

The $JcAIC$ is the measure of the K-L divergence of a model candidate from the true model on the original y scale. Therefore, model candidates can be compared with $JcAIC$ despite of their different response variable. A model with the minimum $JcAIC$ is the optimal model with the corresponding optimal transformation parameter.

Using the derived $JcAIC$ for the Box-Cox transformation, we will compare model candidates whose response variables are Box-Cox transformed with different transformation parameters. However, $JcAIC$ can be also derived for other types of transformations, such as a logarithmic or dual-power transformation (Yang, 2006). The $JcAIC$ always measures the divergence of a candidate model from the true model on the original y scale independent of how the response variable of the model is transformed. Therefore, the $JcAIC$ can compare not only model candidates that use the same transformation with different transformation parameters, but also the models with different types of transformations.

1.3.4 Simultaneous selection of optimal transformation and model formula

As a consequence of the previous sections, we propose the following algorithm to simultaneously select the optimal λ of a Box-Cox transformation and the optimal model among several model candidates. As explained above, considering all possible theoretical M model candidates is often not feasible in practice due to the computational burden. Therefore, the usual step-wise algorithms can be applied where the algorithm stops, if no further improvement can be achieved. In the following, we have chosen *backward* elimination as the exemplary model selection direction. The exchange to *forward* or the extension to *forward-backward* are possible without any difficulties and were done for the simulation experiment in Section 1.4 and the application in Section 1.5.

1) Start with the full model including all P possible x-variables in the data. For the start, the full model is set as the optimal model. Estimate $\hat{\lambda}$ based on the full model to initiate the *backward* model selection.

2) For each step $s = 1, ..., S$:

   i) Consider all possible model candidates which exclude an explanatory variable from the previous optimal model.

   ii) Estimate $\hat{\lambda}$ based on the reduced model formulas and transformed y values $\tilde{y} = T_{\hat{\lambda}}(y)$ for each model candidate. Calculate the $JcAIC$ value from Equation (1.13) with the estimated $\hat{\lambda}$ for each candidate.

   iii) Compare all $JcAIC$ values. The model with the smallest $JcAIC$ value is chosen as the new optimal model for the step.

3) Compare the $JcAIC$ value of the new optimal model in step s with the $JcAIC$ value of the previous optimal model in step $s-1$. If the $JcAIC$ value of the new optimal model is smaller than the previous one, step 2) is repeated until there is no further improvement in terms of $JcAIC$ values.

1.4 Model-based simulation experiment

To support our theoretical findings and the proposed framework from the previous section we conduct simulation studies, which include several settings. The aim of study is to show that under known data settings with a given transformation and model formula, the presented simultaneous algorithm for optimal model and transformation selection depicts the true model for a linear mixed model. The settings include four scenarios: *Normal (1)*, *Normal (2)*, *Log* and *Box-Cox*, each with three explanatory variables. The scenarios are oriented to the simulation study of Rojas-Perilla et al. (2020). The distributions of the explanatory variables are chosen to be representative of both, numeric and categorical variables coded as dummies. The first scenario with normally-distributed random effects and error terms (*Normal (1)*) has an explanatory power of around 40%, and the second (*Normal (2)*) has an explanatory power of 85%, as well as the *Log* and *Box-Cox* scenario. The exact definition of the data settings is given in Table 1.1. In each simulation run (Monte Carlo replication), the explanatory variables, random intercepts and error terms are generated by drawing from the corresponding distributions. Thus, a new pseudo population is created in each simulation run. A total of 500 Monte Carlo replications are generated for each setting. Each of the finite populations consists of $N = 10000$ units evenly divided into $D = 50$ clusters. Within each cluster, a simple random sample is drawn. The cluster-specific sample sizes range from 0 to 29, so that the total sample size sums up to $n = 565$. The distribution of y_{ij} of one population is shown in the Appendix in Table A1 and Figure A1. In addition to the explanatory variables $x_{1,ij}$, $x_{2,ij}$

Table 1.1: Overview of data settings, $i = 1, ..., d, \quad j = 1, ..., N_i$

Data setting	y_{ij}	$x_{1,ij}$	μ_i	$x_{2,ij}$	$x_{3,ij}$	u_i	e_{ij}
Normal (1)	$400 - 10x_{1,ij} + 100x_{2,ij} - 10x_{3,ij} + u_i + e_{ij}$	$\mathcal{N}(\mu_i, 3^2)$	$\mathcal{U}[-3, 3]$	Bin$(1, 0.8)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 30^2)$	$\mathcal{N}(0, 60^2)$
Normal (2)	$400 - 10x_{1,ij} + 100x_{2,ij} - 10x_{3,ij} + u_i + e_{ij}$	$\mathcal{N}(\mu_i, 3^2)$	$\mathcal{U}[-3, 3]$	Bin$(1, 0.8)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 10^2)$	$\mathcal{N}(0, 20^2)$
Log	$exp(10 - x_{1,ij} + x_{2,ij} - 0.5x_{3,ij} + u_i + e_{ij})$	$\mathcal{N}(\mu_i, 2^2)$	$\mathcal{U}[2, 3]$	Bin$(1, 0.8)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 0.4^2)$	$\mathcal{N}(0, 0.8^2)$
Box-Cox	$[(10 - x_{1,ij} + x_{2,ij} - 0.5x_{3,ij} + u_i + e_{ij})(-0.5) + 1]^{\frac{1}{-0.5}}$	$\mathcal{N}(\mu_i, 2^2)$	$\mathcal{U}[2, 3]$	Bin$(1, 0.8)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 0.4^2)$	$\mathcal{N}(0, 0.8^2)$

and $x_{3,ij}$, the random intercepts u_i and the error terms e_{ij}, an additional variable $z_{ij} \sim N(1, 0.1^2)$ is generated in each Monte Carlo replications, which is used to estimate the linear mixed model (but not included in the true data generating mechanism):

$$T_\lambda(y_{ij}) = \widetilde{y}_{ij} = \beta_0 + \beta_1 x_{1,ij} + \beta_2 x_{2,ij} + \beta_3 x_{3,ij} + \beta_4 z_{ij} + u_i + e_{ij}, \tag{1.14}$$

where T denotes the Box-Cox transformation defined in Equation (1.3). In each simulation run, the model selection is performed with four approaches, where the dependent variable y is on different scales:

- on the original scale (no transformation), so that $T_\lambda(y_{ij}) = y_{ij}$ (denoted by *Original*),
- on the log scale, so that $T_\lambda(y_{ij}) = log(y_{ij} + s)$ ($\lambda = 0$) (denoted by *Log*),
- on the Box-Cox scale, so that $T(y_{ij}) = \frac{(y_{ij}+s)^\lambda - 1}{\lambda}$ for $\lambda \in [-2, 2]$ and $\lambda \neq 0$; $log(y_i)$ for $\lambda = 0$ (denoted by *Box-Cox Opt*),

where s denotes the shift parameter $s = |\min(y)| + 1$ only when $\min(y)$ is a negative number. A naive approach which is typically used in applications is

- to perform the model selection on the original scale and afterwards estimate the optimal λ for a Box-Cox transformation (denoted by *Box-Cox Naive*).

In the *Box-Cox Opt* approach the optimal model and the optimal transformation parameter λ are determined simultaneously as described in Section 1.3.4. For each setting, the linear mixed model from the Equation (1.14) and a null model without covariates are estimated. The model selection is than performed with a step-wise algorithm using backward and forward directions based on the $cAIC$ or

$JcAIC$. For the *Original* approach (which operates on the untransformed scale), the $cAIC$ is calculated and the $JcAIC$ in Equation (1.13) is calculated for the other approaches (which operate on the transformed scale). They can be directly compared, as $cAIC$ equals the $JcAIC$ for the *Original* approach. As analytic approximations of the AIC can exhibit negative bias for small sample sizes (Marhuenda et al., 2014), we also use bootstrap versions to estimate the bias correction in the $JcAIC/cAIC$ when a log transformation or no transformation is used. This ensures a fair comparison in the simulation experiment with the estimated $JcAIC$ for a Box-Cox transformation. The bootstrap algorithms to estimate the $cAIC$ for the *Original* and the $JcAIC$ for the *Log* approach are described in the Appendix. The bootstrap algorithms were executed with $B = 200$ replications. In the following, we always refer to $JcAIC$, as in the case of no transformation the $cAIC$ equals the $JcAIC$.

There are three points of interest in the simulation: First, choice of the correct approach for the model selection, second, choice of the transformation parameter and third, choice of the correct transformation and correct model specification. To begin with, we want to evaluate whether the model with the correct approach based on the $JcAIC$ is chosen in agreement with the data setting. For this, we look at the calculated $JcAIC$ values and in relation to this, we also check whether in the case of the Box-Cox transformation the correct associated λ is estimated. Then, we focus on the proportion of simulation runs where the correct transformation is selected and the proportion of correctly specified model formula.

Table 1.2: Summary statistics of *JcAIC* over 500 Monte Carlo replications

Data setting	Approach	Min	1Q	Median	Mean	3Q	Max
Normal (1)	*Original*	6275	6418	6478	6484	6543	6790
	Box-Cox Opt	6271	6418	6478	6484	6543	6786
	Box-Cox Naive	6271	6418	6478	6484	6543	6786
Normal (2)	*Original*	5046	5185	5245	5245	5299	5559
	Box-Cox Opt	5047	5184	5244	5246	5299	5562
	Box-Cox Naive	5047	5184	5244	5246	5299	5562
Log	*Log*	10350	10802	10907	10917	11036	11542
	Box-Cox Opt	10351	10802	10905	10917	11037	11543
	Box-Cox Naive	10435	10878	11001	10993	11101	11726
Box-Cox	*Original*	-296	2603	4497	4821	6434	19141
	Log	-1909	-1597	-1439	-1436	-1301	-792
	Box-Cox Opt	-2572	-2056	-1973	-1969	-1882	-1500
	Box-Cox Naive	-2280	-1961	-1866	-1866	-1775	-971

The parameter λ of the Box-Cox transformation is estimated with the REML algorithm and the simulation is implemented in the statistical programming language R (R Core Team, 2021). For each combination of data settings and approaches the calculated $JcAIC$ are compared and the model with the minimal $JcAIC$ is chosen as optimal. Table 1.2 contains summary statistics of the $JcAIC$ values over the 500 Monte Carlo replications. We observe that in the *Normal (1)* and *Normal (2)* data settings, the calculated $JcAIC$ values of the model with no transformation (*Original*), the Box-Cox transformation (*Box-Cox Opt*), and the *Box-Cox Naive* approach are very close. Often, the calculated $JcAIC$ values for *Box-Cox Opt* and *Box-Cox Naive* are identical. This makes sense considering the corresponding estimated λs in Table 1.3, which are very close to one for both approaches and the resulting distribution close to normality. The deviations of the estimated parameters from one can be explained by the finite population sample from the normal distribution. Looking at the *Log* data setting we see that the distributions of the $JcAIC$ values using the *Log* and the *Box-Cox Opt* approach are very close to each other. Again this makes sense as the estimated λs (see Table 1.3) are close to zero, which results in a log trans-

Table 1.3: Summary statistics of optimal transformation parameter $\hat{\lambda}$ over 500 Monte Carlo replications

Data setting	Approach	Min	1Q	Median	Mean	3Q	Max
Normal (1)	*Box-Cox Opt*	0.4980	0.8800	0.9780	0.9810	1.0970	1.3940
	Box-Cox Naive	0.4980	0.8800	0.9760	0.9810	1.0960	1.3890
Normal (2)	*Box-Cox Opt*	0.4500	0.8830	0.9890	0.9940	1.1140	1.5620
	Box-Cox Naive	0.4500	0.8830	0.9890	0.9940	1.1140	1.5620
Log	*Box-Cox Opt*	-0.0319	-0.0060	-0.0004	-0.0006	0.0051	0.0230
	Box-Cox Naive	-0.0312	-0.0029	0.0037	0.0034	0.0098	0.0309
Box-Cox	*Box-Cox Opt*	-0.5600	-0.4930	-0.4810	-0.4790	-0.4660	-0.3890
	Box-Cox Naive	-0.5510	-0.4940	-0.4810	-0.4790	-0.4650	-0.4070

formation of the data. The $JcAIC$ values of the *Box-Cox Naive* approach are slightly higher. In the case of the *Box-Cox* data setting, the $JcAIC$ values of the *Box-Cox Opt* approach are the smallest, followed by the *Box-Cox Naive* approach. Again, the corresponding estimated λs match the true λ of -0.5 in this case. The values of the *Log* and *Original* approach are considerably higher, which is reasonable given the underlying distribution of the data in this setting. In each setting, the magnitudes and ordering of the values correspond to the underlying distributions of the data and thus to our expectations.

Table 1.4: Proportions [%] of approaches and formulas selected as optimal

Data setting	*Original*	*Log*	*Box-Cox Opt*	*Box-Cox Naive*	*Box-Cox Opt & Naive*	$x1 + x2$	$x1 + x2 + x3$	$x1 + x2 + x3 + z1$	other
Normal (1)	64.1		0.8	0.0	35.1	24.4	60.8	12.0	2.8
Normal (2)	68.9		0.2	0.0	30.9	1.4	86.1	12.5	0.0
Log		70.9	23.2	0.0	5.8	0.6	83.0	14.3	2.1
Box-Cox	0.0	0.0	83.6	0.0	16.4	0.2	81.3	17.3	1.2

Table 1.4 shows the proportions of selected optimal approaches/ transformations and model formulas. For each data setting, the model with the transformation underlying in the data-generating process is selected mostly as optimal, i.e., has the smallest $JcAIC$ values. In the two *Normal* data settings the calculated $JcAIC$ are in around 64% and 69% the smallest, when no transformation is used (*Original*), therefore it is chosen as optimal. This corresponds to the underlying data generating process. In the other cases, *Box-Cox Opt* and *Box-Cox Naive* are chosen as optimal with identical $JcAIC$ values and also very similar estimated λ's, when looking at Table (1.3). This makes sense due to the underlying normal distribution. For normal data, it should make no difference whether the optimal model formula without a transformation is chosen first and then an optimal λ close to one is estimated, or whether the model formula and transformation parameter are chosen simultaneously, as in the *Box-Cox Opt* approach. In the *Log* setting in 71% out of the 500 samples (simulation runs) the true underlying transformation (Log) is chosen as optimal. While in mostly the rest of the samples, the *Box-Cox Opt* approach with optimal $\hat{\lambda}$ near zero, which corresponds to a log transformation, outperforms the *Box-Cox Naive* approach. In 5.8% $JcAIC$ values are identical for both Box-Cox approaches. The advantage of the *Box-Cox Opt* approach is further illustrated in the *Box-Cox* data setting, where this approach is outperforming the other approaches in 83.6% of cases. Looking at the second part of Table (1.4), it can be seen that in settings with high explanatory power (*Normal (2)*, *Log*, *Box-Cox*), the correct model formula ($x_1 + x_2 + x_3$) is selected in over 85% of the simulation runs. However, in the *Normal (1)* setting with lower explanatory power in 60.8% of the samples the correct model formula is selected. This result seems justifiable since, if the explanatory power of the underlying true model is lower, it is more difficult to identify the true underlying relationship. The results emphasize that the presented approach allows for the selection of

the optimal transformation parameter for the Box-Cox transformation and detects the true transformation. In addition, it enables the selection of the correct model formula, whereby the degree depends on the explanatory power of the underlying model.

1.5 Case study: poverty and inequality in municipalities of Guerrero

In this section, the proposed selection approach is applied to data from the state Guerrero in Mexico for estimating poverty and inequality indicators at municipal level. To provide reliable estimates of these indicators at the municipal level, it is necessary to use small area estimation. In order to demonstrate the proposed selection approach, we use a particular small area method - the empirical best predictor (EBP) by Molina and Rao (2010) - which is based on a linear mixed model. In Section 1.5.1, we provide a brief overview of the small area estimation and the EBP. In Section 1.5.2, we describe the data and the problem of simultaneously finding the optimal (linear mixed) model and the transformation parameter. We apply our proposed selection approach and two naive approaches and present the results of the indicators in Section 1.5.3.

1.5.1 Small area estimation and the empirical best predictor

Many surveys are designed to study total populations. For a sample of the total population, direct estimators, for instance the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) can provide reliable estimates due to enough observations/units in the sample. However, direct estimation methods are appropriate only with a sufficient sample size for every domain/area of interest, which is often not the case on a disaggregated regional level. Furthermore, estimators cannot be calculated for domains with no sample data (i.e., out-of-sample domains) or estimators have too large standard errors for domains with only few sample data (Rao and Molina, 2015). When direct estimation cannot provide adequate precision for a domain of interest because of insufficient data, the domain is defined as small and is called small area/small domain (Rao and Molina, 2015; Tzavidis et al., 2018). One way to improve direct estimates is by small area estimation. Small area methods aim to improve the efficiency of the estimation by combining sample data with data from the census/register based on a model (Rao and Yu, 1994; Jiang and Lahiri, 2006). The census/register contains auxiliary variables that may be correlated with the dependent variable and may be used to improve the direct estimates. This is a more complex task as it depends on model building and diagnostics. The model building may include the use of transformation, the selection of the covariates or non-normal error terms.

Since there is no proper survey data which can produce reliable direct estimates of poverty and inequality indicators at municipal level in Guerrero, we use the EBP approach. The approach uses the nested error linear regression model by Battese et al. (1988). This model is a special linear mixed model which includes only random (area specific) intercepts. In the following, we briefly introduce the EBP. Further details are available in Molina and Rao (2010) and Rojas-Perilla et al. (2020).

Assume a finite population of size N divided into D domains. N_i denotes the size of the i-th domain for $i = 1, \cdots, D$. Let y be the target welfare variable (e.g. income) and y_{ij} is the welfare measure of j-th unit in i-th domain where $j = 1, \cdots, N_i$. The sample data does not include all N units in the population but only a part of the population. The sample has a size of n and this sample can also be divided into D domains. n_i denotes the sample size of the i-th domain and it results in $n = \sum_{i=1}^{D} n_i$. Then, the nested error linear regression model is given by

$$y_{ij} = x_{ij}^T \beta + u_i + \varepsilon_{ij}, \tag{1.15}$$

$$u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2), \varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2),$$

where u_i denotes the area random effects and ε_{ij} denotes the error term. Let $\theta = (\beta, \sigma_u, \sigma_\varepsilon)$ be a vector of model parameters. The EBP approach is shortly outlined as follows:

1. Fit the model using the sample data to obtain $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2)$ and $\hat{u}_i$.
2. For $l = 1, ..., L$, generate

   $$\tilde{\epsilon}_{ij}^{(l)} \sim \mathcal{N}(0, \hat{\sigma}_\varepsilon^2), \tilde{u}_i^{(l)} \sim \mathcal{N}(0, \hat{\sigma}_u^2(1 - \hat{\gamma}_i))$$

   for in sample domains,

   $$\tilde{\epsilon}_{ij}^{(l)} \sim \mathcal{N}(0, \hat{\sigma}_\varepsilon^2), \tilde{u}_i^{(l)} \sim \mathcal{N}(0, \hat{\sigma}_u^2)$$

   for out-of-sample domains, using $\hat{\theta}$ with $\hat{\gamma} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2 / n_i}$.
3. Obtain L pseudo-populations by plugging in the explanatory variables in the auxiliary data (i.e. x_{ij}) with $\hat{\beta}$, $\hat{u}_i$, $\tilde{u}_i$ and $\tilde{\varepsilon}_{ij}$ obtained in previous steps into the following model

   $$y_{ij}^{(l)} = x_{ij}^T \hat{\beta} + \hat{u}_i + \tilde{u}_i^{(l)} + \tilde{\varepsilon}_{ij}^{(l)}, l = 1, ..., L$$

   for in sample domains,

   $$y_{ij}^{(l)} = x_{ij}^T \hat{\beta} + \tilde{u}_i^{(l)} + \tilde{\varepsilon}_{ij}^{(l)}, l = 1, ..., L$$

   for out-of-sample domains.
4. Calculate the poverty or inequality indicator for each domain and pseudo population $I_i^{(l)}$, $i = 1, ..., D$ and $l = 1, ..., L$.
5. Take the mean over the L Monte Carlo runs to estimate the EBP of the indicator

   $$\hat{I}_i^{EBP} = \frac{1}{L} \sum_{l=1}^{L} I_i^{(l)}.$$

The EBP with data-driven transformed y is obtained similarly to the described EBP above. The detailed estimation of the EBP with data-driven transformations and corresponding uncertainty measures based on MSE of the EBP are further explained in Rojas-Perilla et al. (2020).

1.5.2 Data and problem

This study uses survey data from the 2010 *Encuesta Nacional de Ingresos y Gastos de los Hogares* (*ENIGH* - National Survey of Household Income and Expenditure) as sample data. This survey is performed every two years by the *Instituto Nacional de Estadística y Geografía* (*INEGI* - The National Institute of Statistics and Geography) and contains socio-demographic information of households, which are also the units of data. *INEGI* also performs the national population and housing census every ten years. As auxiliary data, the census 2010 data is used for the further application.

Guerrero is located in Southwestern Mexico and borders the Pacific ocean. The state is divided into 81 municipalities. 40 municipalities are in the survey data and 41 municipalities are not in the sample. Table 1.5 shows a summary of the number of households per domain in the survey and census data. 1801 households are observed in the sample and on average there are 45 observations per domain. The survey and census data contain a large number of socio-demographic variables. The total household per capita

income in MXN (i.e., `ictpc`) is used as the measurement of welfare. As we used the linear mixed model in Equation (1.15) to explain `ictpc`, the Gaussian assumptions of random effects and errors are required. However, the histogram of `ictpc` in Figure 1.1 shows that the distribution of `ictpc` is very right skewed. Therefore, we apply the Box-Cox transformation to the target variable `ictpc`, such that the violation can be corrected/reduced. For the Box-Cox transformation the optimal transformation parameter λ should be found.

Table 1.5: Number of households per domain in survey and census data

	Min	1Q	Median	Mean	3Q	Max
Survey	13	19	26	45	38	582
Census	585	901	1118	1925	2372	7629

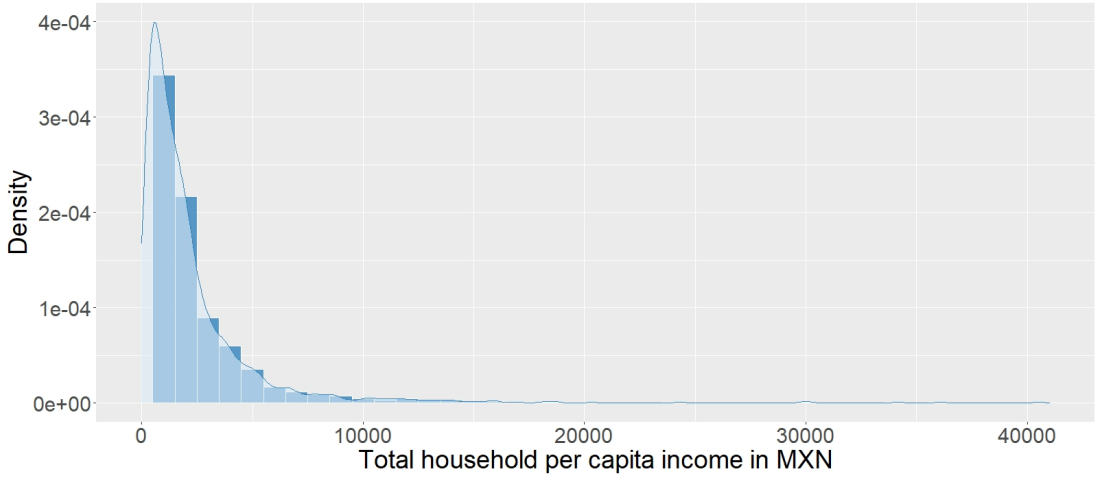


Figure 1.1: Distribution of the total household per capita income in MXN (`ictpc`)

In the survey data there are 34 possible explanatory variables after excluding variables which are highly/perfectly correlated with other variables. We do not know which variables should be included to optimally explain the response variable, therefore, a variable selection should be performed. Consequently, we have two problems to solve: obtaining the optimal transformation parameter λ and finding the optimal model. To solve these problems simultaneously, the optimal transformation parameters are estimated by the REML approach for each model candidate and all model candidates with their own transformed data are compared with the $JcAIC$ introduced in Section 1.3. There are 34 possible explanatory variables in the data, therefore, we theoretically have 2^{34} model candidates. However, fitting these models is unfeasible because of the computational intensity. Instead, a step-wise variable selection proposed in Section 1.3.4 is applied to find the optimal model. With the chosen optimal model the EBP of poverty and inequality indicators are estimated.

To evaluate the EBP based on our optimal model, we apply two naive approaches which are typically used in applications. The first one takes the simple logarithmic transformation to avoid the problem of finding the optimal transformation and performs variable selection based on $cAIC$ on the log-scale. The second specification performs the variable selection initially on the original y scale to find the optimal model. Afterwards, the optimal transformation parameter is chosen based on the optimal model for the original y. Consequently, we have three different EBP estimates: *Box-Cox Opt* denotes the EBP based on our selection method based on the $JcAIC$ as described in Section 1.3, *Log* denotes the first alternative EBP approach and *Box-Cox Naive* denotes the second alternative EBP approach. These three EBP estimates are compared to show that the use of our proposed selection approach based on

the $JcAIC$ can improve the predictive power and reduce the uncertainty of the poverty and inequality estimates.

1.5.3 Results

First, the chosen variables for the optimal model and the optimal transformation parameters of each approach are compared. Table 1.6 shows the chosen variables of each approach and the estimated transformation parameter for models with the Box-Cox transformation. We can see that the results of variable selection can be strongly affected by the response variable. The *Box-Cox Opt* approach performs the variable selection on Box-Cox transformed y and *Log* performs the variable selection on logarithmic transformed y. For these two approaches, a transformation is used to correct the violation of the Gaussian assumptions and then the optimal model is chosen with transformed y. As a result, the chosen variables for the model of *Box-Cox Opt* and *Log* are very similar. In the meantime, the model of *Box-Cox Naive* choose the variables on the original y despite the violation of the Gaussian assumptions in the error terms. As a result, *Box-Cox Naive* has different variables in the model in comparison to the other models. Interestingly, optimal transformation parameters for *Box-Cox Opt* and for *Box-Cox Naive* only differ slightly even though they have many different variables in the models.

Table 1.6: Chosen variables and optimal transformation parameters

EBP approach	Chosen Variables	$\hat{\lambda}$
Box-Cox Opt	$X_1,\ X_2,\ X_3,\ X_4,\ X_5,\ X_6,\ X_7,\ X_8,\ X_9,$ $X_{10},\ X_{11},\ X_{12},\ X_{13},\ X_{14},\ X_{15},\ X_{16},\ X_{17},\ X_{18},\ X_{19},$ $X_{20},\ X_{21},\ X_{22},\ X_{23}$	0.1764
Log	$X_1,\ X_2,\ X_3,\ X_4,\ X_5,\ X_6,\ X_7, X_8, X_9,$ $X_{10},\ X_{11},\ X_{12},\ X_{13},\ X_{14},\ X_{15},\ X_{16},\ X_{17},\ X_{18},\ X_{19},$ $X_{24},\ X_{25}$	-
Box-Cox Naive	$X_1,\ X_2,\ X_3,\ X_4,\ X_5,\ X_6,\ X_7,\ X_8,\ X_9,$ $X_{20},\ X_{21},\ X_{22},\ X_{23},\ X_{26},\ X_{27},\ X_{28}$	0.1888

Second, in order to compare the predictive power of each model, marginal R^2 and conditional R^2 (Nakagawa and Schielzeth, 2013) are calculated and summarized in Table 1.7. The marginal R^2 measures the proportion of variance explained by fixed effects and the conditional R^2 provides the proportion of variance explained by both the fixed and random effects. It is shown that the models with the Box-Cox transformation (i.e., *Box-Cox Opt* and *Box-Cox Naive*) have the higher predictive power than the model with the logarithmic transformation (i.e., *Log*). When we compare *Box-Cox Opt* and *Box-Cox Naive*, we can see that the *Box-Cox Opt*, whose model is optimal for transformed y, has a higher marginal and conditional R^2 than *Box-Cox Naive*, whose model is optimal for the original y scale.

Table 1.7: R^2 of models used for each approach

	marginal R^2	conditional R^2
Box-Cox Opt	0.5997	0.6244
Log	0.5538	0.5878
Box-Cox Naive	0.5630	0.6023

Since the linear mixed model relies on Gaussian assumptions and we decided to use a transformation to correct the violation of the Gaussian assumptions, each approach should be examined concerning whether the violation is corrected. For the examination, the skewness, kurtosis of residuals, and p-

value of the Shapiro-Wilk normality test (Shapiro and Wilk, 1965) on residuals are calculated (Table 1.8). We observe that the logarithmic transformation performs worse than the Box-Cox transformations. For further details we provide quantile-quantile (Q-Q) plots of residuals from the three approaches in Figure A2 in the Appendix. The household level residuals are clearly closer to the normal distribution with transformations. The Box-Cox transformation corrects the violation in household level residuals rather well, however, the residuals slightly deviate in the tails. From the models with the Box-Cox transformation we can at least observe that the municipal level residuals are very close to the normal distribution.

Table 1.8: Skewness, kurtosis and p-value of Shapiro-Wilk test for the household and municipal level residuals

	Household level residuals			Municipal level residuals		
	Skewness	Kurtosis	p-value	Skewness	Kurtosis	p-value
Box-Cox Opt	0.2737	6.3376	0.0000	-0.0893	3.0488	0.7696
Log	-1.4323	13.4906	0.0000	-1.1643	5.9753	0.0089
Box-Cox Naive	0.2329	6.0788	0.0000	-0.0837	3.1332	0.8087

Finally, we want to assess if the improvement in the predictive power of the model due to the proposed simultaneous selection of the transformation and the covariates (*Box-Cox Opt*) translates to more precise small area estimates compared to the two alternative approaches (*Log* and *Box-Cox Naive*). Therefore, we estimate the mean income, head count ratio (HCR) (Foster et al., 1984), and Gini coefficient (Gini, 1912) for the municipalities in Guerrero. To compare the efficiency of these three different approaches, the root mean squared error (RMSE) of the municipal indicators are estimated by a bootstrap (Rojas-Perilla et al., 2020). The RMSE values are visualized in Figure 1.2. Figure 1.2 shows that the *Box-Cox Opt* is the most efficient approach, since for all three indicators it has the smallest estimated RMSE. When the naive approaches are compared, we cannot say which approach is more efficient because for some indicators *Log* has the smaller RMSE and for other indicators *Box-Cox Naive* has the smaller RMSE. It seems that the model and transformation selection is especially important for parameters associated with the tails of the distribution.

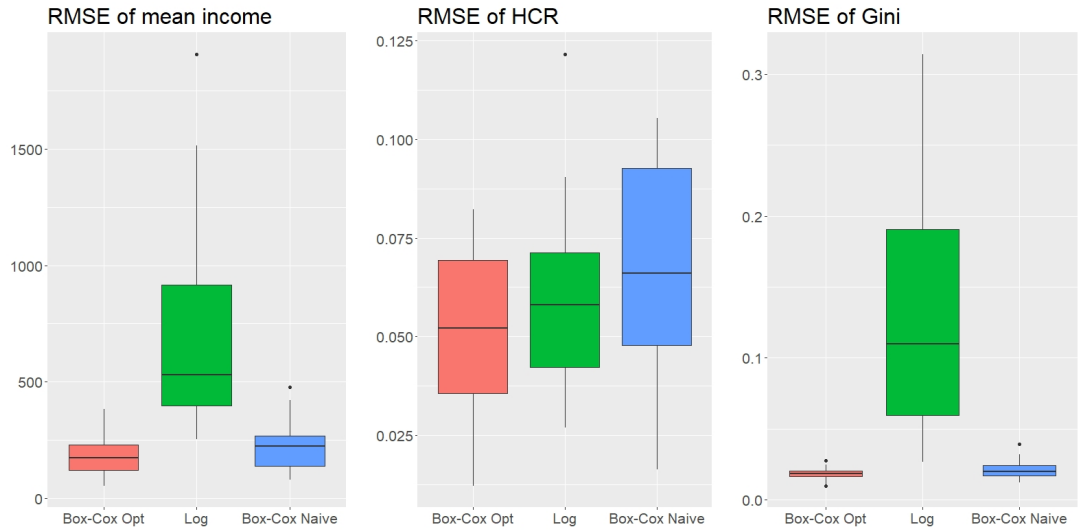


Figure 1.2: RMSE of EBP estimates for mean income, HCR and Gini

Figure 1.3 shows EBP estimates of mean income, HCR, and Gini of municipalities in Guerrero based on *Box-Cox Opt* approach. The Southwestern part of Guerrero, which resembles the coastline

(Costa Grande region and Acapulco), features a tourism industry which contributes to the municipalities having a higher mean income. Furthermore, along a north-south axis between Chilpancingo in the south and Taxco in the north, numerous industries are concentrated. These industries focus on the production of handcrafted items using local resources. This also contributes to a higher income in these municipalities. Consequently, the HCR and Gini coefficient in these municipalities are lower than the others. This means, that the people in these municipalities earn more money than in other municipalities and the wealth is more equally distributed compared to other municipalities. On the other hand, the eastern part of Guerrero is suffering from higher levels of poverty and inequality. Municipalities in the region are covered with mountains and when compared to all other regions of Guerrero, these municipalities exhibit the highest number of indigenous people living there.

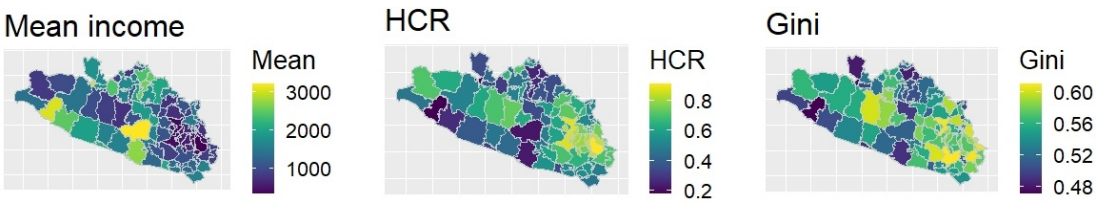


Figure 1.3: EBP estimates for mean income, HCR and Gini based on *Box-Cox Opt* approach

1.6 Conclusions and future research directions

The main purpose of this study was to find a solution to two practical problems in the context of linear mixed models: a) the true model for explaining the response variable is unknown and b) the model assumptions, especially the Gaussian assumptions of the error terms, are violated. While these problems commonly appear together, we provide a solution to find the optimal model and the optimal transformation simultaneously. We focus on one of the most commonly used transformations, the Box-Cox transformation. Since the $cAIC$ is scale dependent, we provide an adjusted $cAIC$ by using the Jacobian of the transformation such that different models with differently compared transformed response variables can be compared. As a large number of possible explanatory variables increases computational costs, we propose an optimal simultaneous selection approach based on Jacobian adjusted $cAIC$ ($JcAIC$), which is also feasible for a large number of variables. Our model-based simulation studies show that the proposed selection approach chooses the true model with a transformation parameter close to the true value in most cases and performs better compared to naive selection approaches. The proposed simultaneous selection approach can be used in many different areas of research. As an example, we provide a case study where we apply the selection approach for estimating poverty and inequality indicators at municipal level in Mexico. We observe that the model selected by the proposed simultaneous approach has higher predictive power than other approaches. The improvements in terms of predictive power and model building translate to more precise small area estimates of the poverty and inequality indicators.

Further research should be shifted towards alternative variable selection criteria. For instance, Bunke et al. (1999) show that the cross validation selection criterion can simultaneously select the optimal parametric model and the optimal transformation parameter of the Box-Cox transformation for nonlinear regression models. Furthermore, Fang (2011) proves that the $cAIC$ is asymptotically equivalent to the leave-one-observation-out cross validation for linear mixed models. Therefore, deriving the cross validation selection criterion for the linear mixed model and comparing the results with the $JcAIC$ might be a promising avenue for further research. The selection based on cross validation criterion may improve the quality of the prediction. Moreover, it is also possible to derive the $JcAIC$ for other transformations which require the estimation of the transformation parameter. The use of $JcAIC$

as a selection criterion between different transformations with different optimal models is also a potential research direction. However, it should be noted that the use of our proposed approach is less useful when the point of interest is to interpret the effect of the chosen explanatory variables on the original scaled data. Gurka et al. (2006) introduce a bias corrected beta coefficient for linear mixed models under the Box-Cox transformation which produces a more precise interpretation of the beta coefficients. However, the interpretation does only hold for the transformed response variable. On the original scaled response, it is not clear how strong the effects of the explanatory variables are. To enable interpreting the effects of explanatory variables on the original data, further research is needed for general regression models with the Box-Cox transformed response variable.

Declarations

The authors did not receive support from any organization for the submitted work.

Acknowledgments

Lee gratefully acknowledges support by a scholarship of Studienstiftung des deutschen Volkes. The authors are grateful for the computation time provided by the HPC service of the Freie Universität Berlin. Finally, the authors are indebted to the Joint Editor, Associate Editor and two referees for comments that significantly improved the paper.

Appendix A

A.1 Bootstrap for *Original* and *Log* approach

Bootstrap for *Original*

1. Fit the model in Equation (1.1) to obtain estimates of model parameters $\hat{\theta}$.
2. Generate $u^{(b)}$ from $\mathcal{N}(0, \hat{G}_0)$ and $\varepsilon^{(b)}$ from $\mathcal{N}(0, \hat{\sigma}^2)$ and create bootstrap y using

$$y^{(b)} = X\hat{\beta} + Zu^{(b)} + \varepsilon^{(b)}.$$

3. Re-fit the model with the bootstrap sample $y^{(b)}$ and obtain bootstrap estimates of model parameters $\hat{\theta}^{(b)}$ and $\hat{u}^{(b)}$.
4. Obtain the BC by

$$\begin{aligned} BC =& \frac{1}{B}\sum_{b=1}^{B}\left[-\frac{1}{2\hat{\sigma}^{2(b)}}\left(y^{(b)} - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\right)^T\left(y^{(b)} - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\right)\right] \\ &+ \frac{1}{B}\sum_{b=1}^{B}\left[\frac{1}{2\hat{\sigma}^{2(b)}}\left(y - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\right)^T\left(y - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\right)\right]. \end{aligned}$$

Bootstrap for *Log*

1. Transform the y to the $\widetilde{y}$ using $\widetilde{y} = \log(y + s)$.
2. Fit the model with $\widetilde{y}$ to obtain estimates of model parameters $\hat{\theta}$.
3. Generate $u^{(b)}$ from $\mathcal{N}(0, \hat{G}_0)$ and $\varepsilon^{(b)}$ from $\mathcal{N}(0, \hat{\sigma}^2)$ and create bootstrap $\widetilde{y}$ using

$$\widetilde{y}^{(b)} = X\hat{\beta} + Zu^{(b)} + \varepsilon^{(b)}.$$

4. Re-fit the model with bootstrap sample $\widetilde{y}^{(b)}$ and obtain bootstrap estimates of model parameters $\hat{\theta}^{(b)}$ and $\hat{u}^{(b)}$.

5. Back-transform $\widetilde{y}^{(b)}$ to obtain $y^{(b)}$. $y^{(b)}$ is obtained by $y^{(b)} = \exp(\widetilde{y}^{(b)}) - s$.

6. Obtain the BC by

$$BC = \frac{1}{B}\sum_{b=1}^{B}\left[-\frac{N}{2}log\big(2\pi\hat{\sigma}^{2(b)}\big) - \frac{1}{2\hat{\sigma}^{2(b)}} \cdot \left(\widetilde{y}^{(b)} - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\right)^T \left(\widetilde{y}^{(b)} - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\right) + J(y^{(b)})\right] - \frac{1}{B}\sum_{b=1}^{B}\left[-\frac{N}{2}log\big(2\pi\hat{\sigma}^{2(b)}\big) - \frac{1}{2\hat{\sigma}^{2(b)}}\left(\widetilde{y} - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\right)^T \left(\widetilde{y} - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\right) + J(y)\right],$$

with

$$J(y^{(b)}) = -\sum_{i=1}^{D}\sum_{j=1}^{N_i}\log(y^{(b)} + s), \qquad J(y) = -\sum_{i=1}^{D}\sum_{j=1}^{N_i}\log(y + s).$$

A.2 Graphics and tables

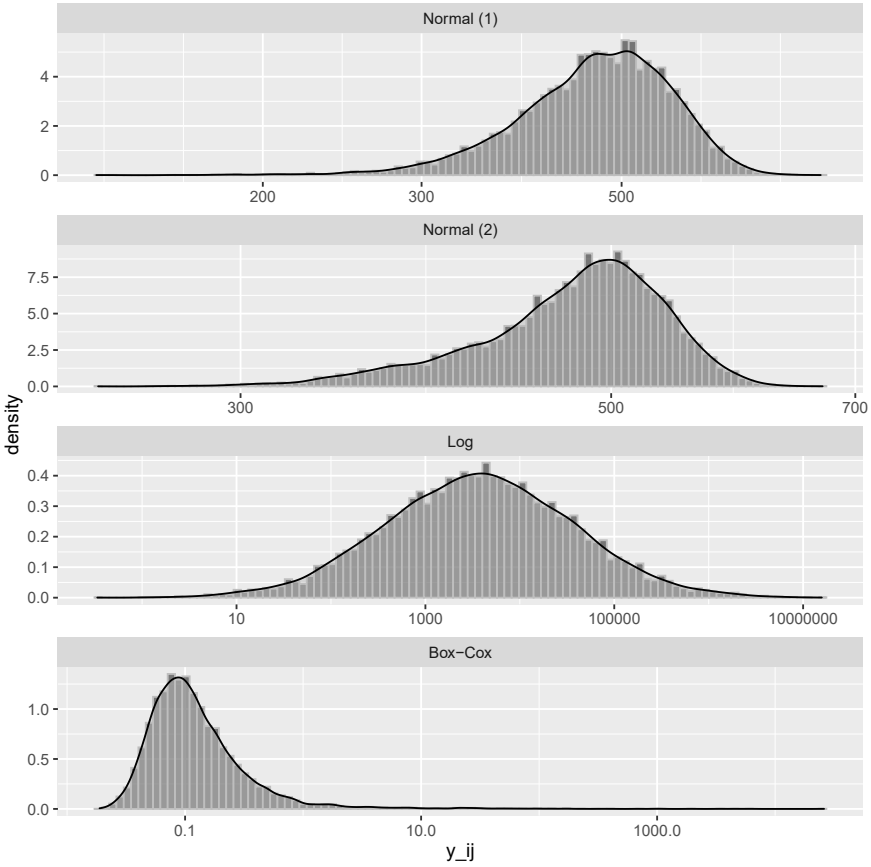


Figure A1: Density of the dependent variable (y_{ij}) in the first Monte Carlo population. Note that a base-10 log scale is used for the x-axis for the *Log* and *Box-Cox* setting

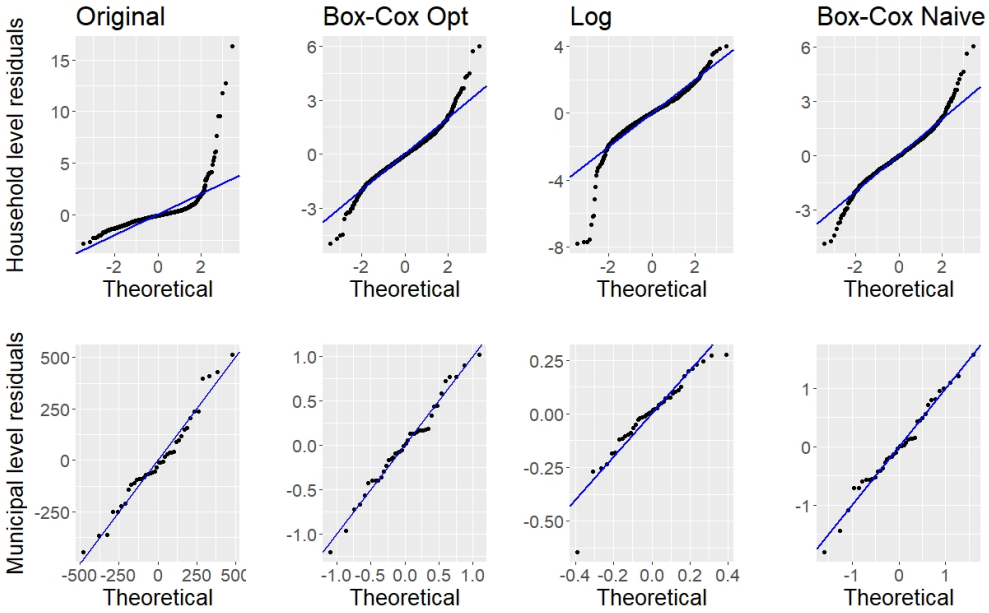


Figure A2: Q-Q plots for household level and municipal level residuals of different EBP approaches

Table A1: Summary statistics of the dependent variable (y_{ij}) in the first Monte Carlo population

Data setting	Min	1.Q	Median	Mean	3.Q	Max
Normal (1)	131	416	476	475	535	831
Normal (2)	247	442	484	477	517	669
Log	0	793	3732	48861	17384	15769695
Box-Cox	0.019	0.066	0.103	5.064	0.183	25978.438

Chapter 2

Estimation of the consumer price index with regional weights using small area estimation methods: A case study of Germany

Abstract

The consumer price index (CPI) is an important indicator for formulating effective policies related to wage and inflation control. Most countries regularly produce the national CPI and some countries also publish the CPI at sub-national levels. This sub-national (or regional) CPI depicts region specific information and consequently is a helpful tool for local policy makers. Germany provides this regional CPI for all 16 states with national product weights. Still, national products weights do not adequately represent the importance of products at the regional level, whereas a better regional CPI uses regional product weights to reflect regional specifics more accurately. In this study, I explore the estimation of such a better regional CPI with regional weights by using an accessible income and consumption survey from Germany. To obtain reliable regional weights, I focus on estimating regional expenditures for each product. Regional weights of each product are derived by calculating the proportion of the estimated specific product expenditure over the total expenditure. However, estimating regional expenditures is challenging because of the small sample size in each region. Small sample sizes potentially produce unreliable estimates. To address this problem, I propose the use of a small area estimation approach based on multivariate Fay-Herriot (MFH) models. By using MFH models, I show how model-based estimation of regional expenditure improves the reliability of the estimation using the case of Germany and further discuss the limitations as well as future research directions.

Keywords: inflation, regional inequality, regional difference, price statistics, household consumption

2.1 Introduction

The consumer price index (CPI) is one of the most important economic indicators and vital for formulating policy. It measures the change in price of the household shopping basket of consumed goods and services. Therefore, the estimation of the CPI is essential, especially for formulating inflation policies. Consequently, national statistical offices in most countries compute the national CPI every month and use it mainly for measuring inflation as well as support for directing monetary policies. Inflation is

usually understood to be similar within a monetary union such as the eurozone or a country. However, many empirical studies show a heterogeneity of regional inflation levels within a monetary union and even between the regions in a single country (Duarte and Wolman, 2008; Nagayasu, 2011; Weinand and von Auer, 2020). This means that an estimation of a reliable sub-national CPI (i.e., regional CPI) is important for fully understanding regional economic heterogeneity. Despite the clear need for a regional CPI, regular publications of regional CPI by national statistical offices are not yet common because of the associated difficulties.

Publishing a reliable regional CPI is difficult due to a lack of data. To produce a reliable regional CPI, we need sufficient information about prices and household expenditure on consumed goods and services at the regional level. Existing survey data usually includes only a small number of observed prices and households at the regional level which makes the estimation of regional CPI unreliable. Consequently, many countries do not publish the regional CPI and only some countries provide the regional CPI. Additionally, each country decides independently on which regional level to include. For example, the German Federal Statistical Office (FSO) provides the regional CPI for 16 states (Statistisches Bundesamt, 2021b), whereas the U.S. Bureau of Labor Statistics (BLS) publishes regional CPI monthly for 4 regions, 9 divisions, and bimonthly for 23 metro areas (BLS, 2023b) with regional prices but national product weights. The national statistical office of Canada regularly provides regional CPI for 10 provinces and for capital cities of 3 territories, also without regional specific product weights (Statistics Canada, 2023). The Statistics Bureau of Japan (SBJ) computes the regional CPI with area specific product weights, but only for the special wards of Tokyo (SBJ, 2022). The use of national product weights for the regional CPI is not appropriate seeing that national product weights do not adequately represent the importance of the goods and services in each region (Kosfeld et al., 2009). However, when we use regional product weights derived from limited household expenditure data at the regional level, the resulting regional weights may be unreliable (BLS, 2023a).

To obtain reliable regional product weights, we must reliably estimate regional household expenditure of consumed goods and services first. Small area estimation approaches are especially useful for this purpose. SAE methods enable the estimation of target indicators which are not observed in some regions and generally improve the reliability of estimates based on a model for small regions. Researchers have applied SAE methods in connection with regional households expenditure before with different focuses: Marchetti and Secondi (2017) suggest the estimation of the regional household consumption expenditure using the Fay-Herriot model (Fay and Herriot, 1979) to construct purchasing power parities at the provincial level in Italy. However, they only estimate the regional total household expenditure, and not the expenditure on each consumed product. Pusponegoro et al. (2021) model the regional CPI in Indonesia with the Fay-Herriot model combined with a penalty for random effects. The dependent variable of the model in the study is the regional CPI of each product category. For each category, they use an univariate Fay-Herriot model. Dawber et al. (2022) estimate region specific expenditures on consumed goods and services for 12 regions in UK based on the Fay-Herriot model combined with a smoothing method. By using estimated regional expenditures, they estimate regional specific product weights. The approaches of Pusponegoro et al. (2021) and Dawber et al. (2022) use the univariate Fay-Herriot model, which means that they assume that the expenditure on a product or the regional CPI on each expenditure category does not affect the expenditure or the CPI on other products. This, however, is an unrealistic assumption.

In the present study, I focus on the reliable estimation of the regional CPI for 16 states in Germany. FSO regularly publishes the regional CPI for 16 states but these values are computed with the national weights. There is enough data for regional prices but not for regional product weights. Therefore, the main purpose of this study is reliably estimating regional expenditures of products which are used to compute the regional weights. The related studies use univariate Fay-Herriot models to estimate expenditure of products. This study, however, suggests to use multivariate Fay-Herriot models, as the

univariate model automatically assumes that expenditures on each good and service are independent of other goods and services. This is a doubtful assumption because the prices of some goods and services affect the demand of relevant goods and services such as substitutes or complementary goods. For example, when people buy bread, they also tend to buy butter. This means that the expenditure on goods and services cannot be completely independent from each other. Therefore, I use multivariate Fay-Herriot (MFH) models (Benavent and Morales, 2016) which allows the correlation between the expenditures on different products. Based on this model, the expenditures on each product are accurately estimated at the state level of Germany. The estimated regional expenditures are used to obtain the regional product weights which constructs the reliable regional CPI.

I will proceed by defining CPI in Section 2.2 and datasets in Section 2.3. In Section 2.4, I introduce and compare the direct estimation method and the MFH method for the regional households expenditure on each product. Using the most reliable estimation results from Section 2.4, I construct the regional CPI at German state level in Section 2.5. Finally, in Section 2.6, I will discuss relevant limitations and challenges as well as further research directions.

2.2 Consumer price index

The CPI measures the change of the household shopping basket price between the target period t and the base period 0. The shopping basket includes goods and services consumed by households. Most countries (including Germany) use the classification of individual consumption by purpose (COICOP) to classify the goods and services for CPI calculation. FSO uses the German COICOP by Statistisches Bundesamt (2013). Table 2.1 shows an overview and examples of German COICOP categories at different COICOP levels. The price of the basket is defined as the sum of the prices of the goods and services in the basket multiplied by its consumed quantity at a chosen COICOP level as follows

$$\sum_{k=1}^{K} p_k \cdot q_k, \tag{2.1}$$

where p_k is the average price of the kth category/product of a chosen COICOP level and q_k is the consumed quantity of the kth category/product at a chosen COICOP level. Raw prices of a product are elementally aggregated at the deepest COICOP level according to the Dutot index (Dutot, 1738) in Germany. The Dutot index is furthermore aggregated to obtain the p_k at other aggregated COICOP levels.

The CPI consists of two baskets: a basket at the target period t and a basket at the base period 0. There are two different types of CPI: Laspeyres type and Paasche type. The prices of each basket are estimated at each period but the quantities are fixed at base period 0 for the Laspeyres type index, while Paasche type index fix the quantities at target period t. Germany uses the Laspeyres type index to obtain national CPI in target period t (Statistisches Bundesamt, 2021b). This type of index is defined as follows

$$CPI^t = \frac{\sum_{k=1}^{K} p_k^t q_k^0}{\sum_{k=1}^{K} p_k^0 q_k^0} \cdot 100,$$

where p_k^0 is price of product k in base period 0, p_k^t is price of product k in period t, and q_k^0 is the consumed quantity of kth product in base period 0. To obtain the average prices at each COICOP category (i.e., p_k), the price data is required. Research Data Center of the Federal Statistical Office and Statistical Offices of Federal States (RDC) of Germany regularly collects the prices of goods and services consumed by private households (RDC, 2021). The raw prices of the items are collected from the outlets including both physical and online shops. These item prices are aggregated at a chosen COICOP level so that there is one average price for each COICOP category. Further details of the

aggregation process are explained in Section 2.3.2. However, how often each product is consumed is commonly not observed. Therefore, the national CPI in Germany is calculated by using the transformed Laspeyres index given by

$$CPI^t = 100 \cdot \sum_{k=1}^{K} \frac{p_k^t}{p_k^0} w_k^0. \tag{2.2}$$

w_k^0 is the proportion of expenditure on kth product in base period 0 and estimated by using the information about the household expenditure. These proportions are used as product weights and can be understood as the importance of each product.

Table 2.1: COICOP by Statistisches Bundesamt (2013)

Level	COICOP 2	COICOP 3	COICOP 4	COICOP 5	COICOP 6	COICOP 7
Label	Division *(Abteilungen)*	Group *(Gruppen)*	Class *(Klassen)*	Sub-class *(Unterklassen)*	Category *(Kategorien)*	Sub-category *(Unterkategorien)*
Example	Food and non-alcoholic beverages	Food	Meat and meat products	Dried, salted, smoked and other meat and sausage products	Other sausage	Boiled sausage (incl. Bratwurst)
Code of Example	01	011	0112	0112 7	0112 72	0112 721

The regional CPI is constructed similarly to the national CPI in Equation (2.2). The regional CPI from FSO uses the regional prices with the national product weights as follows

$$CPI_i^t = 100 \cdot \sum_{k=1}^{K} \frac{p_{ik}^t}{p_{ik}^0} w_k^0, \tag{2.3}$$

where the index i denotes the state of Germany and p_{ik} is the price of product k in state i. This study aims to estimate the regional CPI with the regional product weights defined as

$$CPI_i^t = 100 \cdot \sum_{k=1}^{K} \frac{p_{ik}^t}{p_{ik}^0} w_{ik}^0, \tag{2.4}$$

where w_{ik}^0 is the regional product weight for kth product in state i in the base period. By using the regional weight, the regional CPI can better represent the importance of products in each region.

2.3 Data sources

For the CPI estimation, the FSO uses multiple data sources. The product weights are estimated every five years based on *Die Einkommens- und Verbrauchsstichprobe* (*EVS* - Income and consumption survey) data with supplementary economic account data (Statistisches Bundesamt, 2021b). The prices are collected by *Verbraucherpreisindex für Deutschland* (*VPI* - CPI for Germany) data (Statistisches Bundesamt, 2021b). As the latest *EVS* is from 2018, I use this data (*EVS 2018*) and the corresponding *VPI* data from 2018 (*VPI 2018*). These datasets are provided by Research Data Center of the RDC. Unfortunately, economic account data is not fully accessible. Therefore, the regional CPI is estimated with these given data sources.

2.3.1 EVS 2018 (RDC, 2018a)

The *EVS* is carried out every five years to provide an overview of the social and socio-economic status of private households in Germany. Around 60000 households are asked to participate in the survey. The households are selected through quota sampling, using the latest results of the German microcensus. *EVS* consists of four survey parts: a) *Allgemeine Angaben* (*AA* - general information), b) *Geld- und Sachvermögen* (*GS* - monetary and tangible assets), c) *Haushaltsbuch* (*HB* - household budget book), and d) *Feinaufzeichnungsheft für Nahrungsmittel, Getränke und Tabakwaren* (*NGT* - fine recorded budget book for food, beverages, and tobacco products). The participating households fill in the questionnaires for general information including the socio-demographic and socio-economic status of the household and of the household members (*AA*) and the questionnaires of monetary and tangible assets (*GS*). Furthermore, the households record their income and consumption for three months (*HB*) and 20% of participants are asked to record their consumption of food, beverages, and tobacco products at COICOP 7 level in detail for a month (*NGT*). Based on the collected data, the *EVS* provides information about the income situation and the consumption pattern of the households in Germany. Therefore, CPI uses the *EVS* to estimate the product weights.

For this study, I use *EVS 2018*. The RDC provides collected data divided into six files: a file for each survey part (i.e., one file for *AA*, *GS*, *HB*, *NGT*), a combined file with survey parts except for *NGT*, and a file providing personal information of the households. To estimate the product weights at the deepest COICOP level of the *EVS* data (i.e., COICOP 7 level), I need the consumption information for each product at this level. Then, I aggregate this information for the estimation of the weights at different COICOP levels. Although the file for *NGT* provides detailed consumption information at COICOP 7 level, unfortunately, there is no file which provides detailed consumption information for all COICOP 7 products. As a consequence, I concentrate on estimating product weights (and CPI) for food, beverages, and tobacco products.

NGT data from *EVS 2018* includes consumption information of 10351 households in 235 COICOP 7 categories. Table 2.2 shows the number of households per state in *NGT* data with the total number of households in each state based on census data (Statistisches Bundesamt, 2018). The state codes are given in Appendix (Table B1). On average, 0.28% of all households in each state are observed in *NGT*. The data also includes two grossing-up factors. As *EVS* uses quota sampling, there is no information about the design weights or inclusion probabilities. Instead, grossing-up factors are given so that we can aggregate the data without bias either at state level or at country level (Statistisches Bundesamt, 2022). The direct estimates of product weights are obtainable by combining consumption variables and grossing-up factors. Moreover, the data provides some socio-demographic information of households such as average age of household members, education level, etc. We can, therefore, build small area models with the direct estimates and these socio-demographic variables.

Table 2.2: Number of households per state in *NGT* data and in the census data (Statistisches Bundesamt, 2018)

	SH	HH	NI	HB	NW	HE	RP	BW	BY	SL	BE	BB	MV	SN	ST	TH
NGT	401	258	904	122	1953	938	529	1136	1512	121	393	322	317	722	413	310
Census (in 1000)	1470	1003	3973	366	8756	3091	1961	5286	6453	493	2028	1257	830	2156	1151	1104

2.3.2 VPI 2018 (RDC, 2018b)

The consumer price statistics collect the prices of domestically consumed goods and services by private households. The prices of items are collected every month either locally through the statistical offices of the federal states or centrally by the FSO. Centrally collected prices are the same for the 16 states

in Germany. As the prices are collected in different quantity and measurement (e.g., 1 orange vs. 1 kg oranges), they are adjusted according to quantity, measurement, etc. The statistical offices of the federal states perform the elementary aggregation with the adjusted raw prices according the Dutot index so that each COICOP 10 category has one elementary price index (RDC, 2021). The description of COICOP 10 categories is only available in the *VPI* data, although it is not explicitly defined in Statistisches Bundesamt (2013). *VPI* data is published every year and includes different monthly prices (e.g., raw item prices, adjusted prices, and elementary price indices) with further information such as shop type, state, shop type weights, state weights, COICOP 10 code, and the national product weights from the previous base year 2015 at COICOP 10 level.

I use *VPI 2018* to keep the time consistent with the *EVS* data. *VPI 2018* includes the prices of 190 COICOP 10 food, beverages, and tobacco products. The data includes monthly prices of products from 8 different shop types in 16 different states. Each COICOP 10 product has one elementary price index per month, state, and shop type. Assume $p^m_{ik_{10}s}$ denotes the elementary price index of the kth COICOP 10 product from the sth shop type in the mth month in state i. For the regional CPI, we need the product price per state for the target period t and for the base period 0. The target period in this paper is each month of 2018. Therefore, each product price of each month per state must be calculated to construct the CPI. For the target month t in 2018, the price is aggregated with the shop type weights as follows

$$p^t_{ik_{10}} = \sum_{s=1}^{8} p^t_{ik_{10}s} \cdot SW_s,$$

where SW_s denotes the shop type weight of the sth shop type. Moreover, the prices per state for the base period are also required to construct the CPI. As the base period is a whole year, the monthly prices are equally weighted to obtain the price per state for the year as follows

$$p^0_{ik_{10}} = \frac{1}{12} \sum_{m=1}^{12} p^m_{ik_{10}}.$$

2.4 Estimation of regional product expenditure

The product weights for CPI are the proportions of the expenditures on each consumed good and service. Germany changes the base period every five years and the product weights are calculated accordingly every five years. The *EVS* is a critical component for the estimation of national product weights in Germany at a detailed COICOP level because the FSO calculates the product expenditure using the *EVS* and updates the results with additional data, since the *EVS* takes place two years prior to the base year (Statistisches Bundesamt, 2023). As the updating method with supplementing data is not given, the regional expenditure are estimated by focusing on *EVS* data. In this paper, I consider two possible ways for the product weight estimation: a) direct estimates and b) model-based estimates. I explain the estimation methods in this section and estimate the total expenditure of each product using both methods from COICOP 7 to COICOP 3 level. The estimations are implemented by using the statistical programming language R (R Core Team, 2022) with R packages **emdi** (Kreutzmann et al., 2019; Harmening et al., 2023) and **msae** (Permatasari and Ubaidillah, 2021). I furthermore evaluate and compare the estimation results to find the optimal level and method for deriving the product weights for the regional CPI.

2.4.1 Direct estimates

Let y_{ijk} denote the total expenditure for the kth product at a certain COICOP level of the jth household in the ith region and w_{ik} denote the product weights of kth product in ith region. w_{ik} is the proportion

of the expenditure on the kth COICOP product from the total expenditure in the ith state. Let θ_{ik} denote the expenditure on the kth COICOP product and $\sum_{k=1}^{K} \theta_{ik}$ denote the total expenditure in the ith state. The direct estimates of the total expenditures are obtained by using the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952). This requires inclusion probabilities, however, *EVS* does not have inclusion probabilities because it uses a non-probability sampling. Instead, *EVS* provides two grossing-up factors which can replace the sampling weights for computing an unbiased HT estimator (Statistisches Bundesamt, 2021a). Then, the direct estimate of the total expenditure of the kth product at national level ($\widehat{\theta}_k^{direct}$) is computed as follows

$$\widehat{\theta}_k^{direct} = \sum_{i=1}^{D} \sum_{j=1}^{n_i} y_{ijk} \cdot GF_{ij}^{national},$$

where n_i is the number of households in the EVS data for the ith region and $GF_{ij}^{national}$ denotes the grossing-up factor of the jth household in the ith region for the national level indicator calculation. Using $\widehat{\theta}_k^{direct}$, the direct estimates of the national product weights are derived. The FSO uses these national product weights to provide the regional CPI of each state. However, I am looking to calculate regional weights. For this, I need to estimate regional total expenditures. As the *EVS* includes the grossing-up factor for the state level aggregation, the direct estimates of the regional indicators can be obtained in the same way

$$\widehat{\theta}_{ik}^{direct} = \sum_{j=1}^{n_i} y_{ijk} \cdot GF_{ij}^{state}, \tag{2.5}$$

where GF_{ij}^{state} denotes the grossing-up factor for the state level estimation.

To evaluate the reliability of estimates between different COICOP levels, I calculate the coefficient of variation (CV) values of the estimated total expenditures of products per state from COICOP 7 to COICOP 3 level. CV is given by $\sqrt{\widehat{V}(\widehat{\theta}_{ik}^{0,direct})}/\widehat{\theta}_{ik}^{0,direct}$. The variance of the estimators (i.e., $\widehat{V}(\widehat{\theta}_{ik}^{0,direct})$) are obtained for the CV calculation by using the naive bootstrap method (Alfons and Templ, 2013). Figure 2.1 shows the distributions of the CV per state at COICOP 5, COICOP 4, and COICOP 3 levels. We can see that the CV values become much smaller at COICOP 4 level compared to COICOP 5 level. The CV values at COICOP 3 level are similar to COICOP 4 level. The proportion of the estimators which have a CV value higher than 0.2 is 34.67% at COICOP 5, 7.08% at COICOP 4, and 7.81% at COICOP 3. Therefore, it is reasonable to estimate the regional product weights at COICOP 4.

2.4.2 Model-based estimates

The direct estimates generally show good results, although there is still a reliability problem and missing products because, if the number of observations in a state is not high enough, the expenditure of infrequently consumed product is rarely observed or completely missing. Therefore, the expenditures of the unobserved products cannot be directly estimated and the estimation of some states' products are not reliable. SAE methods are a solution to these problems. Combining direct estimates of the expenditures with auxiliary data based on a small area model enables the estimation of unobserved products. Moreover, an accurate small area model improves the reliability of the estimation.

SAE methods combine survey data and auxiliary information based on a model to provide more precise estimates. They can be classified into two groups according to the type of auxiliary data: unit-level models and area-level models. While area-level models require auxiliary information at area level, unit level models require auxiliary data with all units in the population of interest such as census data. In this study, the Fay-Herriot (FH) model (Fay and Herriot, 1979) is appropriate because the auxiliary

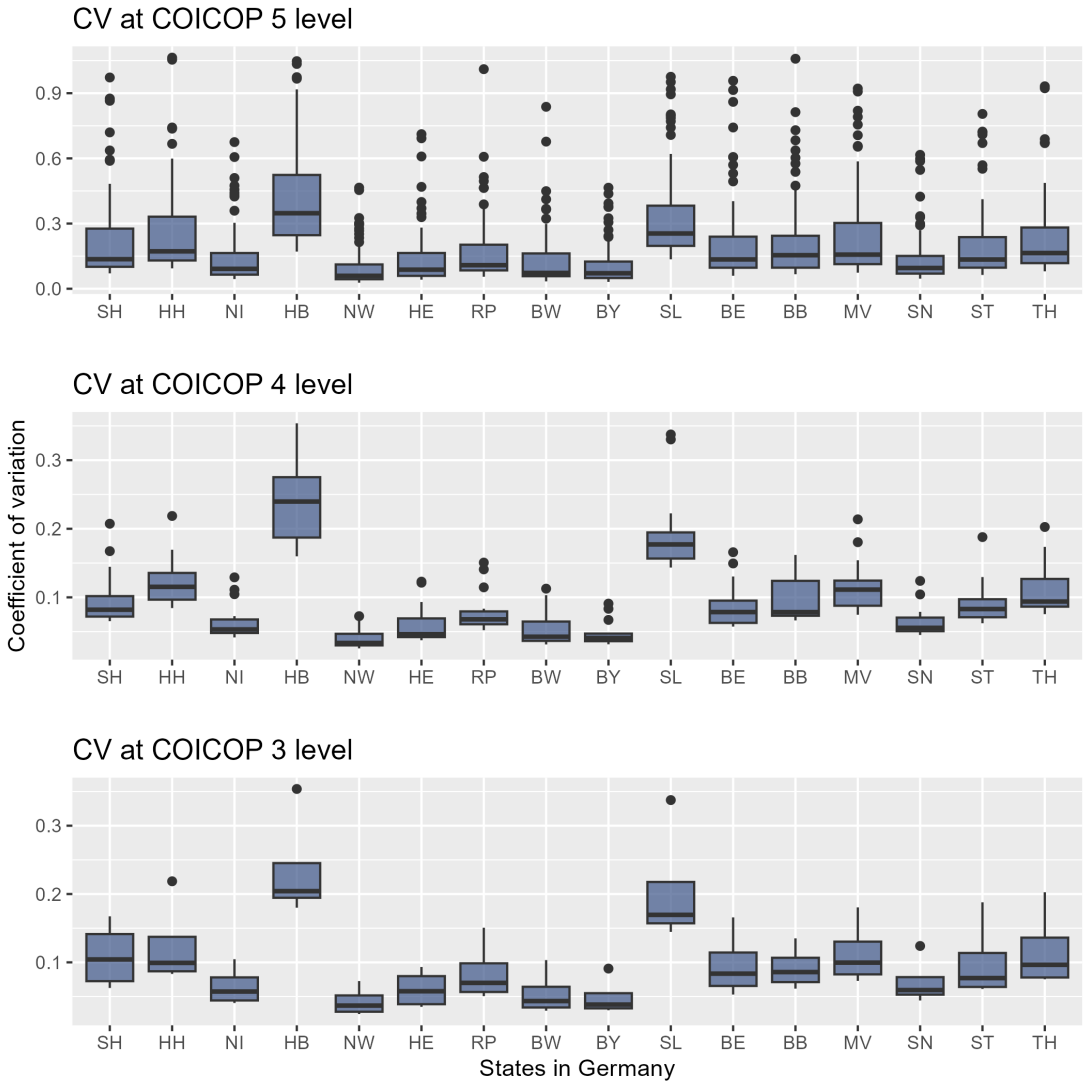


Figure 2.1: CV of $\hat{\theta}_{ik}^{direct}$ at different COICOP levels in each state of Germany

data can be obtained at area level. The total expenditures of products per state are estimated for each COICOP product using a FH model. Since the total expenditures per state must be estimated for all products and they are potentially correlated, it is reasonable to use a MFH model.

Benavent and Morales (2016) introduce MFH models for a finite population divided into D regions. $\theta_i = (\theta_{i1}, \cdots, \theta_{iK})^T$ is a vector of target indicators which are the total expenditure of products at a certain COICOP level in region i. e_i denotes the sampling errors which are normally distributed with a covariance matrix V_{e_i}. V_{e_i} corresponds to the variances and covariances of the $\widehat{\theta}_i^{direct}$ assumed to be known. The model assumes

$$\widehat{\theta}_i^{direct} = \theta_i + e_i,\ e_i \sim \mathcal{N}(0, V_{e_i})$$

and a linear relation between target indicators and explanatory variables. $x_{ik} = (x_{ik1}, \cdots, x_{ikp_k})$ denotes the explanatory variables for the kth indicator in the ith region, where p_k is the number of explanatory variables related to the kth indicator. Let $x_i = diag(x_{i1}, \cdots, x_{ik}, \cdots, x_{iK})$ be the matrix of explanatory variables and u_i be the independent and normally distributed random effects. Then, the

MFH model for the ith region is given as

$$\widehat{\theta}_i^{direct} = x_i\beta + u_i + e_i$$

with $u_i \overset{ind}{\sim} \mathcal{N}(0, V_{u_i})$. The MFH model can be written in matrix form as follows

$$\widehat{\theta}^{direct} = X\beta + Zu + e,\ e \sim \mathcal{N}(0, V_e),\ u \sim \mathcal{N}(0, V_u) \tag{2.6}$$

with the design matrix for the fixed effects X and for the random effects Z. Benavent and Morales (2016) introduce four MFH models which differ in the structures of V_e and V_u. For the model-based estimation, the first two models defined as Model 0 and Model 1 by Benavent and Morales (2016) are used. Model 0 assumes $V_{e_i} = diag(\sigma^2_{e_{ik}})$ and $V_{u_i} = diag(\sigma^2_{u_{ik}})$. This means that e_d and u_d are independent and that indicators are independent from each other. Therefore, Model 0 is equivalent to the univariate FH model. Using Model 1, V_{u_i} is unchanged but it allows a non-diagonal matrix for V_{e_i}, which means Model 1 takes into account the relation between the indicators within the regions. V_{e_i} is defined as follows

$$V_{e_i} = \begin{pmatrix} \sigma^2_{i1} & \sigma_{i,(1,2)} & \cdots & \sigma_{i,(1,K)} \\ \sigma_{i,(2,1)} & \sigma^2_{i2} & \cdots & \sigma_{i,(2,K)} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{i,(K,1)} & \sigma_{i,(K,2)} & \cdots & \sigma^2_{iK} \end{pmatrix}, \tag{2.7}$$

where $\sigma_{i,(k,l)}$ denotes the covariance between the kth and lth components in the ith region for $k, l = 1, \cdots, K$. The MFH estimator of the target indicators is obtained for both models by

$$\widehat{\theta}^{MFH} = X\hat{\beta} + Z^T\widehat{V}_u Z\widehat{\Omega}^{-1}(y - X\hat{\beta}), \tag{2.8}$$

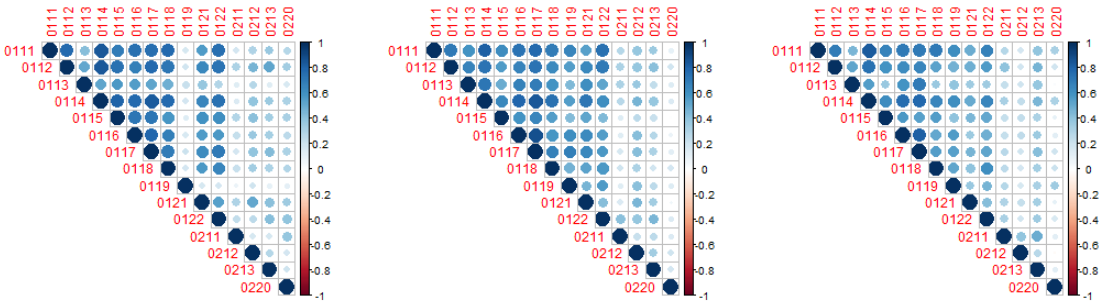
where $\widehat{\Omega} = Z^T\widehat{V}_u Z + V_e$ and $\hat{\beta} = (X^T\widehat{\Omega}^{-1}X)^{-1}X^T\widehat{\Omega}^{-1}y$. For further details about the estimation of the model parameters, see Benavent and Morales (2016) and Permatasari and Ubaidillah (2021).

When the indicators are independent from each other, Model 0 is a reasonable choice. However, as explained in the introduction, it is much more sensible to assume a potential relation between expenditures on certain products. Therefore, Model 1 is a better choice for the estimation since it takes into account the relations of indicators in the covariance matrix of the error term. Whether this relation actually exists is examined by the Pearson's correlation coefficients between direct estimates of total expenditures in each state at different COICOP levels. Table 2.3 shows the summary of median correlation coefficients between direct estimates in each region at different COICOP levels. From COICOP 7 to COICOP 5, most products have very low correlations with other products. At lower COICOP levels, median correlation values in each region between the products exceed 0.3. Furthermore, Figure 2.2 shows that each state has different correlation patterns between products at COICOP 4 level. The description of COICOP 4 codes is given in Appendix (Table B2). For example, the COICOP 4 code 0119 (which denotes other food products) generally has a very small correlation with other products at COICOP 4 level in Brandenburg compared to Hamburg and Berlin. The data shows that, at some COICOP levels, the correlation between direct estimates of total expenditures cannot be ignored. That is why it is reasonable to take the following into account: Model 1 is used for the estimation of the total expenditures at COICOP 4 and COICOP 3 levels, while Model 0 is used at other COICOP levels.

Before fitting the model, the model assumptions must be examined and the model specification done. The MFH models assume normally distributed errors. However, income or expenditure type variables usually cannot fulfill this assumption. The direct estimated total expenditures of nearly all products in *EVS 2018* show very right skewed distributions which is evidence of the normality assumption vio-

Table 2.3: Median correlation coefficients between estimated regional total expenditures of products at different COICOP levels

	Min	1.Q	Median	Mean	3.Q	Max
COICOP 7	0.0694	0.0822	0.0922	0.0924	0.1041	0.1131
COICOP 6	0.0937	0.1072	0.1204	0.1206	0.1295	0.1734
COICOP 5	0.1174	0.1374	0.1428	0.1488	0.1638	0.1996
COICOP 4	0.3230	0.3996	0.4372	0.4433	0.4864	0.5704
COICOP 3	0.3218	0.3644	0.3947	0.4146	0.4893	0.5450



(a) Brandenburg (b) Hamburg (c) Berlin

Figure 2.2: Correlation coefficients between estimated regional total expenditures of COICOP 4 products in (a) Brandenburg, (b) Hamburg, and (c) Berlin

lation. Therefore, in this study, when some products expenditures show an evidence of the violation, the logarithmic transformation is applied to these response variables to correct the violation. For those indicators, the model in Equation (2.6) is fitted with $\log(\widehat{\theta}^{direct})$ instead of $\widehat{\theta}^{direct}$. Moreover, the model assumes the know V_e. Consequently, the variances for direct estimators and the covariances between direct estimators are estimated by a naive bootstrap method (Alfons and Templ, 2013) and used in V_e. For the model specification, socio-demographic variables are extracted from the *EVS 2018* at state level so that they can be used as x variables in the MFH model. As the model is fitted at state level, there are only 16 observations per indicator. Therefore, only one explanatory variable that has the highest correlation with each response variable is applied to the model. This means that the number of explanatory variables related to the kth indicator equals to 1 for all indicators (i.e., $p_k = 1,\ \forall k$). By fitting the model, the MFH estimator of θ is obtained for transformed products on the logarithmic scale. As the MFH estimates on the original scale are needed, the crude type back-transformation (Harmening et al., 2023) is performed for these products as suggested in Rao and Molina (2015) and Neves et al. (2013).

To evaluate the reliability of the MFH estimates, CV values of $\widehat{\theta}^{MFH}$ are calculated at each COICOP level. Table 2.4 shows the median CV value in each state in Germany. The MSE of the estimators are obtained for the CV calculation by using the MSE estimation method by Benavent and Morales (2016) and Permatasari and Ubaidillah (2021). As the COICOP levels become more aggregated, the estimates become more reliable with COICOP 4 being the most reliable. Below COICOP 4, however, this no longer occurs. For the comparison between direct and model-based estimation, $\widehat{\theta}_{ik}^{direct}$ in Equation (2.5) and $\widehat{\theta}_{ik}^{MFH}$ in Equation (2.8) are estimated at COICOP 4 level and the CV of the estimates are calculated. Figure 2.3 shows that the direct estimates of total expenditures are generally reliable with CV values under 0.2. MFH estimation shows an improvement in most states compared to the direct estimates at COICOP 4 level and the use of model-based estimation is necessary to obtain reliable estimates for some states like Bremen (HB).

Table 2.4: Median CV of MFH estimators for total expenditures of products per state and for Germany (GER) at different COICOP levels

	SH	HH	NI	HB	NW	HE	RP	BW	BY	SL	BE	BB	MV	SN	ST	TH	GER
COICOP 7	0.18	0.22	0.12	0.36	0.09	0.12	0.15	0.11	0.10	0.31	0.18	0.19	0.21	0.13	0.18	0.21	0.200
COICOP 6	0.15	0.20	0.10	0.33	0.07	0.10	0.12	0.08	0.09	0.26	0.15	0.16	0.19	0.11	0.15	0.18	0.151
COICOP 5	0.13	0.17	0.09	0.31	0.06	0.09	0.11	0.07	0.07	0.24	0.13	0.15	0.15	0.09	0.13	0.16	0.135
COICOP 4	0.08	0.11	0.05	0.18	0.03	0.05	0.07	0.04	0.04	0.14	0.08	0.08	0.11	0.05	0.08	0.09	0.076
COICOP 3	0.10	0.10	0.06	0.20	0.04	0.06	0.07	0.04	0.04	0.16	0.08	0.09	0.10	0.06	0.08	0.10	0.079

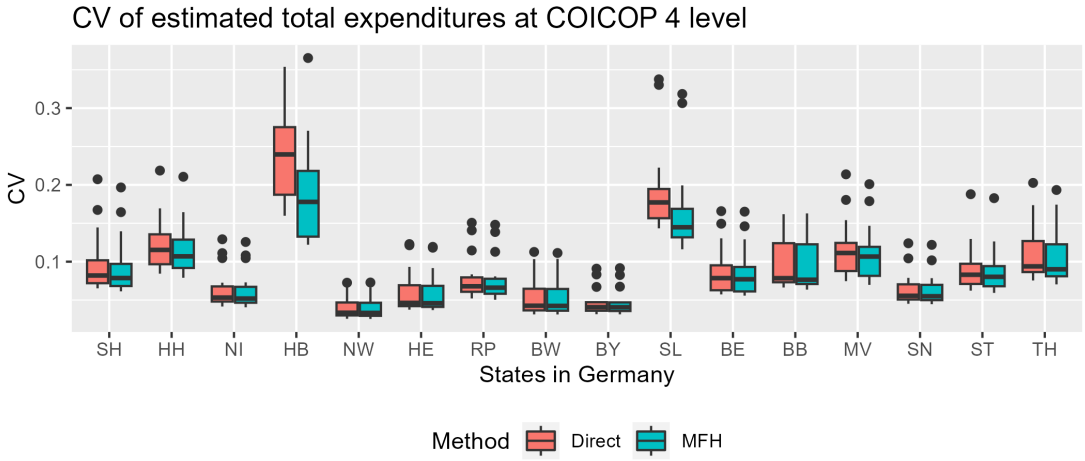


Figure 2.3: CV of direct and FH estimators at COICOP 4 level

2.5 Estimation of regional CPI

As MFH estimates at COICOP 4 level are the most reliable, the CPI is constructed at COICOP 4 level. However, *VPI* data provides prices only at COICOP 10 level. The prices should be aggregated to COICOP 4 level. For the price aggregation, the product weights of the next COICOP level are used. For example, under the COICOP 6 product 0116 16 (berries and grapes), there are three COICOP 7 products 0116 161 (strawberries), 0116 165 (grapes), and 0116 169 (other berries). Let $p^t_{ik_6}$ be the price of the kth product at COICOP 6 level (e.g., price for the product 0116 16) in the ith region, $p^t_{ic_7}$ denotes prices of COICOP 7 products which belongs to the kth COICOP 6 product (e.g., prices of products 0116 161, 0116 165, 0116 169). For the price aggregation, product weights are necessary. Let $\widehat{w}_{ic_7}$ be a vector of the corresponding estimated product weights at COICOP 7 level (e.g., product weights of 0116 161, 0116 165, 0116 169) in the ith region. Then, the price of the kth product at COICOP 6 level is defined as the aggregation of weighted COICOP 7 prices of products belonging to the kth COICOP 6 product as follows

$$p^t_{ik_6} = \sum_{c_7 \in k_6} p^t_{ic_7} \cdot \frac{\widehat{w}^0_{ic_7}}{\sum_{c_7 \in k_6} \widehat{w}^0_{ic_7}} \quad \forall k_6. \tag{2.9}$$

The price aggregation is done in the same way as in Equation (2.9) for every other COICOP level. To aggregate the prices to COICOP 4 level, the product weights from COICOP 10, 7, 6, and 5 levels are required. The MFH estimates are used for price aggregation from COICOP 7 to COICOP 5 levels. Since the *EVS* data are not available at COICOP 10 level, the product weights cannot be estimated at this level. Therefore, the given national COICOP 10 level products weights in the *VPI* data are used. For the other

COICOP levels, the product weights are derived by using

$$\widehat{w}_{ik}^{0,MFH} = \frac{\widehat{\theta}_{ik}^{MFH}}{\sum_{k=1}^{K} \widehat{\theta}_{ik}^{MFH}} \text{ for } (k,K) \in \{(k_7,K_7),(k_6,K_6),(k_5,K_5)\}.$$

After the price aggregation, the CPI is estimated using

$$\widehat{CPI}_i^t = 100 \cdot \sum_{k_4=1}^{K_4} \frac{p_{ik_4}^t}{p_{ik_4}^0} \widehat{w}_{ik_4}^{0,MFH}$$

for each state in each month of 2018. The product weights are derived from the estimated regional expenditure on each consumed goods and services in Equation (2.5) as follows

$$\widehat{w}_{ik_4}^{0,MFH} = \frac{\widehat{\theta}_{ik_4}^{MFH}}{\sum_{k=1}^{K_4} \widehat{\theta}_{ik_4}^{MFH}}. \tag{2.10}$$

Furthermore, the percent change of CPI from the previous month to the current month is calculated from February to December for each state in Germany. The percent change of CPI in target month t in state i is defined as

$$\frac{\widehat{CPI}_i^t}{\widehat{CPI}_i^{t-1}} \cdot 100 - 100,$$

where $t-1$ denotes the previous month.

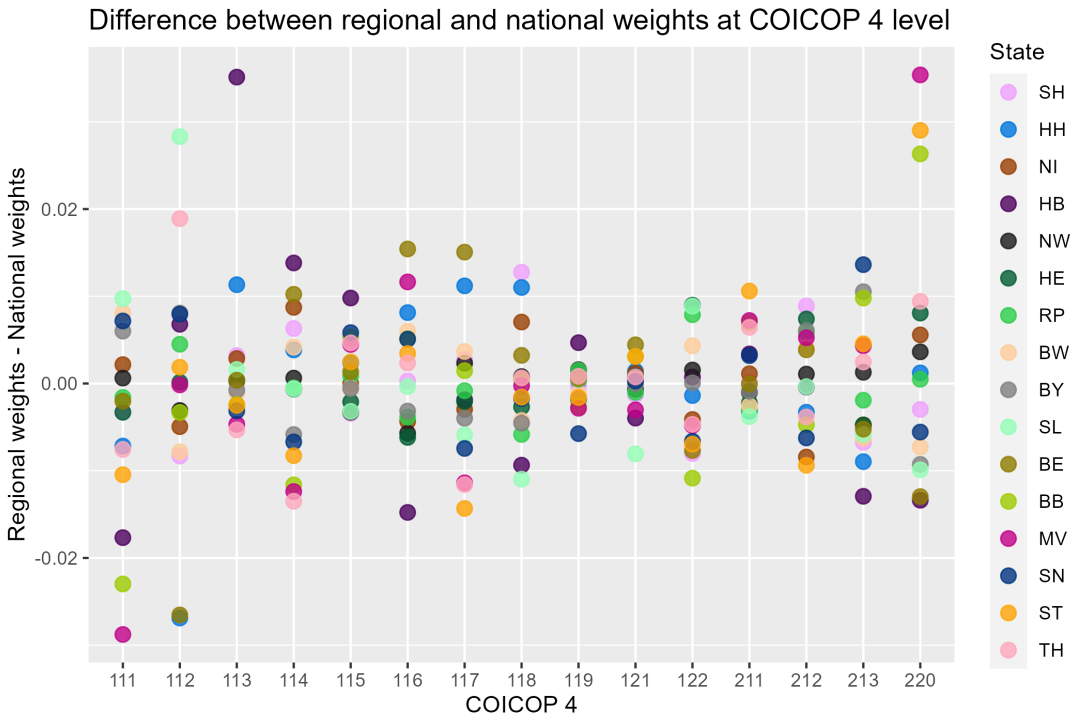


Figure 2.4: Difference between model-based regional weights ($\widehat{w}_{ik}^{0,MFH}$) and direct estimated national weights ($\widehat{w}_{ik}^{0,direct}$) at COICOP 4 level

It is impossible to exactly reproduce the regional CPI with national weights from FSO due to a lack of publicly accessible data. Still, to make a comparison of regional CPI with national and regional

weights possible, the national weights are derived as follows

$$\widehat{w}_{ik_4}^{0,direct} = \frac{\widehat{\theta}_{ik_4}^{direct}}{\sum_{k=1}^{K_4} \widehat{\theta}_{ik_4}^{direct}} \tag{2.11}$$

and used to calculate the CPI in each state. Figure 2.4 shows the difference between regional weights in Equation (2.10) and national weights in Equation (2.11) at COICOP 4 level. Deviations from national weights differ depending on product. Products such as 111 and 220 have a wider range of difference compared to products like 119 and 121. Each state shows a specific consumption pattern. For example, Hamburg (HH) and Bremen (HB) show a higher consumption of fish products (113) whereas Berlin (BE) shows a much lower consumption of meat products (112) together with a higher expenditure on fruits (116) and vegetables (117) compared to the national level.

Figure 2.5 shows the CPI with national product weights on the red line and with regional weights on the blue line for Brandenburg, Hamburg, and Berlin for every month in 2018. The CPI and the changes of CPI in other states are available in Appendix (Figure B1 and Figure B2). In Figure 2.5, we observe that Brandenburg and Berlin have a similar pattern: CPI values are high at the beginning of the year and decrease during the summer and increase again in the winter. Hamburg shows a different pattern compared to the others. Although the patterns of CPI are similar with national weights and regional weights, the exact values of the CPI with regional weights differ from the CPI values with national weights. Berlin shows a higher CPI with regional weights from January to June and higher CPI values with national weights from June to the end of the year. Hamburg shows the opposite, meaning that the CPI with regional weights is higher until May and that it decreases smaller from May on. Unlike these two states, the difference between national and regional weights are not as distinct as in Brandenburg. This suggests that the national weights are only in some states representative of the importance of products and that the type of used weights affects the CPI results.

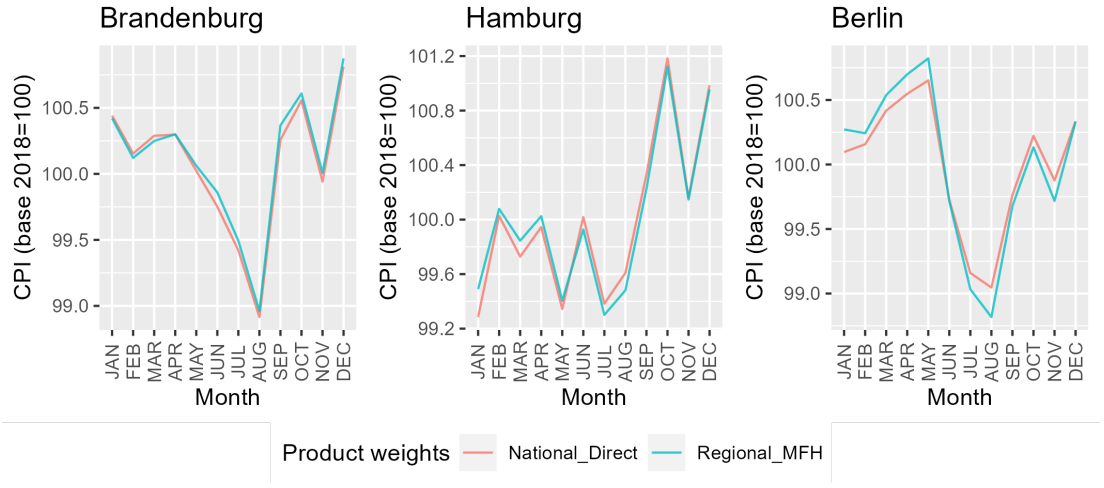


Figure 2.5: Estimation of regional CPI in Brandenburg, Hamburg, and Berlin

Figure 2.6 shows the change in percentage of CPI from the previous month to the current month for the same states in 2018. Despite the differences in CPI with different weights, the changes in CPI between national weights and regional weights are not as clear as in Figure 2.5. A possible reason for this is the pattern of CPI with regional weights not differing that much from the pattern with national weights. Still, it should be noted that only months within the base year are investigated. The difference in CPI and changes in CPI with regional and national weights may increase over time. Furthermore, the presented regional CPI only include food, beverages, and tobacco products. Since the expenditure on these products is not as high as expenditures on housing and transport costs, the expenditure pattern for

housing and transport costs may differ much more, depending on the region. Consequently, the overall regional CPI and changes of CPI including all products with regional weights will deviate more strongly from the regional CPI from the national weights.

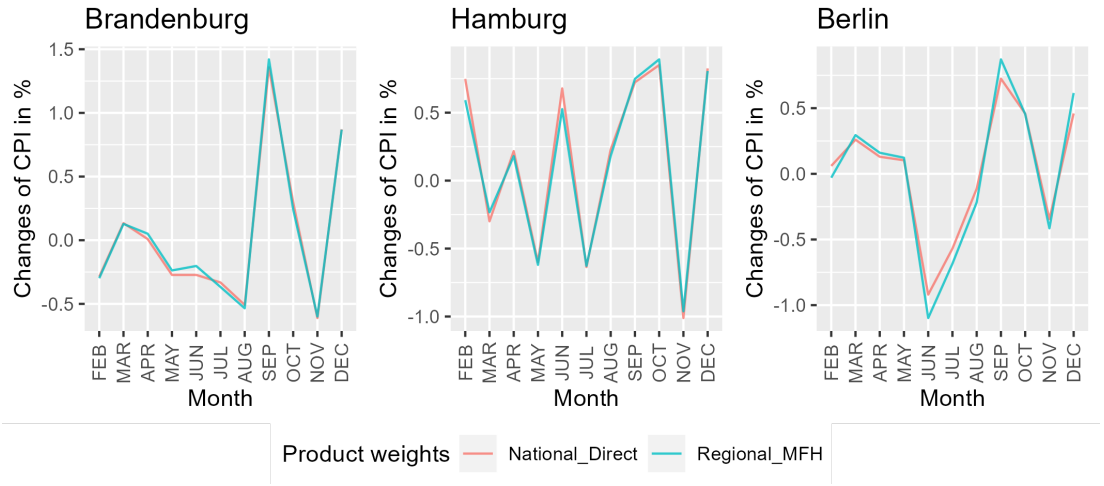


Figure 2.6: Changes of regional CPI from the previous month in % for Brandenburg, Hamburg, and Berlin

2.6 Discussing problems and solutions

In the present study, I propose an estimation method of regional CPI with regional products weights for food, beverages, and tobacco products using existing data sources. For the estimation, I focus on estimating reliable regional product expenditures. As the direct estimation method cannot provide the regional product expenditures for some products because of data deficiencies, I introduce a model-based estimation with a SAE method. This method provides regional product weights for all products and improves the reliability of estimators. Although this study provides a solution for estimating reliable regional product expenditures for the construction of the regional CPI, there are still limitations and challenges that need to be considered such as limited data and possible estimator improvement.

2.6.1 Limited data and information availability

A challenge for SAE research dealing with the regional CPI in Germany is that the consumption information of the households is only partly available in the *EVS* data. To overcome this lack of information, FSO uses additional data sources. However, these data sources are not accessible to the public, which makes the estimation of product weights for all consumed goods and services impossible. Consequently, the regional CPI is estimated only for food, beverages, and tobacco products. Moreover, both data sources (*EVS* and *VPI*) provide a regional identifier for states. With the given data sources, it is very difficult to estimate the regional CPI at a more disaggregated regional level. However, regional CPI at a more disaggregated level such as at district or at municipal level are essential for better understanding problems at the local level and for subsequently formulating effective policies to deal with these problems. The introduced model-based estimation method is applicable (if the data is available) for CPI with all goods and services and at more disaggregated regional levels.

A further problem is that the sampling design for the price data is unknown. Instead, only the elementary price indices at state level are provided in the *VPI* data so that the calculation of regional CPI at state level is possible. As a result, the provided indices are treated as known fixed values, ignoring the uncertainty caused from price estimation or from elementary aggregation. This means that

the uncertainty from price components are not considered during the construction of regional CPI while the reliability of product weights are examined with the CV of MFH estimates. Additionally, the details of product weights estimation processes and price aggregation processes from FSO are not available. Therefore, it is impossible to reproduce the exact results of CPI from FSO.

2.6.2 Further challenges

Model extensions can improve estimation results. For the FH model, the covariates must be measured without error (Fay and Herriot, 1979). However, the auxiliary x variables used in this study are estimated from the *EVS* data. This means that the auxiliary variables could include errors. On that account, an extension of the MFH models with the measurement error FH models (Ybarra and Lohr, 2008; Burgard et al., 2020) could be considered to construct a more efficient model-based estimation. Additionally, a smoothing method with pooled data is a possible extension when the number of regions in the FH model is small. The FH model at German state level has only 16 observations. Because of this small number of observations, challenges such as a restriction in the number of covariates persist. Dawber et al. (2022) deal with similar challenges by combining a smoothing method and a SAE method with pooled data from multiple years. This increases the number of observations in the FH model and improves the temporal instability problem. Moreover, a MFH model can be extended for it to handle the correlation of indicators between regions. The MFH model in this study takes into account the relations of products within regions by allowing a non-diagonal covariance matrix for the errors. But the model assumes a diagonal covariance matrix of random effects, which means the regions are independent from each other. To take into account the correlations between regions, an extension of MFH with a more general covariance matrix of errors can be considered when computing time allows. As another alternative, a spatial FH model (Pratesi and Salvati, 2008) may also be possible.

I estimate the total expenditure of each product using an MFH model and build the MFH estimators of weights based on the total expenditure estimates, which means that additional uncertainty may arise. To solve this problem, alternative models for estimating compositional indicators can be considered so that the product weights can be directly used as the response variable of the model. Scealy and Welsh (2017) propose a mixed model for compositional expenditure data based on the Kent distribution. As Scealy and Welsh (2017) do not consider small area problems, an extension for a SAE context is needed. Esteban et al. (2020) introduce a trivariate FH model to estimate compositions. A generalization from a trivariate to a MFH model is necessary for applying this method to the estimation of product weights. Additionally, Krause et al. (2022) propose a robust MFH model to predict compositions and examine the model with a simulation study and real data application for a trivariate case. They observe that this model solves the problem of measurement errors. However, when the models from Esteban et al. (2020) and Krause et al. (2022) are applied to the product weights estimation, there are numerous unknown model parameters. Therefore, computational costs of estimation must be considered for higher dimensional MFH models.

The estimated CPI cannot be free from sampling errors. The variance of estimated indicators is commonly used as a measurement of uncertainty caused by using sample data. However, estimating variance of CPI is a challenging task and most countries do not provide uncertainty measurements of the CPI. There is no clear consent on how the variance of the estimated CPI should be measured since each country has their own estimation method for the CPI. The CPI has different components and needs multiple estimation techniques with a combination of multiple data sources in the whole estimation process. Additionally, the used data sources are changing as well. These make the estimation of the sampling variance for CPI more difficult. Nonetheless, some studies propose estimation methods for sampling error estimation (see Smith (2021)). For example, the BLS is regularly providing the variance estimates of the CPI based on a jackknife resampling method (BLS, 2023a). The exact variance

estimation method of the BLS might not be adaptable for all countries, although a resampling based approach such as jackknife or bootstrap could generally be a possible solution for complex indices such as CPI.

Recently, many countries experienced dramatic changes in household consumption patterns caused by the aftereffects of COVID-19. Additionally, some countries faced problems collecting price data and consumer expenditure data because of lockdown policies. National statistical offices in most countries could not provide correct information regarding prices and consumption patterns with the CPI during the pandemic for two reasons. First, the product weights can no longer represent the importance of products since the consumption pattern of households changed extremely fast in a short period. Second, there are too many missing prices because of restrictions on, for example, bars or restaurants that could not operate at all. Consequently, each country made its own decision on how to provide the CPI according to their data situation. For example, Australia, Canada, and the UK have been either annually or biennially updating the product weights and needed an update in 2021 based on pre-pandemic data. However, they used partial data from 2021 so that the changed consumption pattern could be taken into account for the revision (United Nations, 2021). In the meantime, Germany revised the weights in 2023 for the new base year 2020 based on data from 2019 to 2021 (Statistisches Bundesamt, 2023). The changed consumption pattern caused by the pandemic could not be taken into account in the CPI until the revision was done. It stands to reason that an annual or biennial cycle of weights revision makes the acquired index much more resilient to outside shocks like war or other disasters with long-term effects. Consequently, states like Canada or the UK produce CPI that represent reality much better than states that only do so every five years. As the FSO already updates the product weights for the harmonized CPI annually, a more frequent revision for CPI should not be too difficult.

2.7 Conclusion

The present study contributes to the growing body of literature on the dynamics between national and regional price levels. More specifically, I emphasize the importance of working with reliable CPI estimates for producing accurate economic analyses, which in turn are indispensable for formulating effective policy recommendations. For this purpose, I construct a reliable estimation method for the regional CPI in Germany. The FSO provides the CPI for Germany and the regional CPI for each state with the same national product weights, which may not accurately represent the importance of the goods and services at the regional level. To accurately depict regional levels, I focus on the estimation of regional product expenditure and propose a regional CPI estimation with regional product weights.

Because most national statistical offices use direct estimation, I first estimate regional weights based on direct estimations. However, since the expenditure for some products is not always observed at the regional level, direct estimates are not always applicable. To alleviate this issue, I adapt a SAE method. Furthermore, as there are numerous correlated products which have to be included for the estimation, I apply MFH models. From COICOP 4 level, the expenditure of products show a correlation between the expenditures. Therefore, I use a MFH model allowing relation between products at COICOP 4 and COICOP 3 levels. By using the model-based estimates, the regional expenditures for all products, including the products which are not observed in the data, can be estimated. Consequently, the regional CPI can be reliably estimated with the regional product weights. While this regional CPI estimation method applies only to Germany here, it can theoretically be adjusted and used by any country to reliably estimate regional expenditures and the regional CPI.

Acknowledgments

Lee gratefully acknowledges support by a scholarship of Studienstiftung des deutschen Volkes.

Appendix B

B.1 Graphics and tables

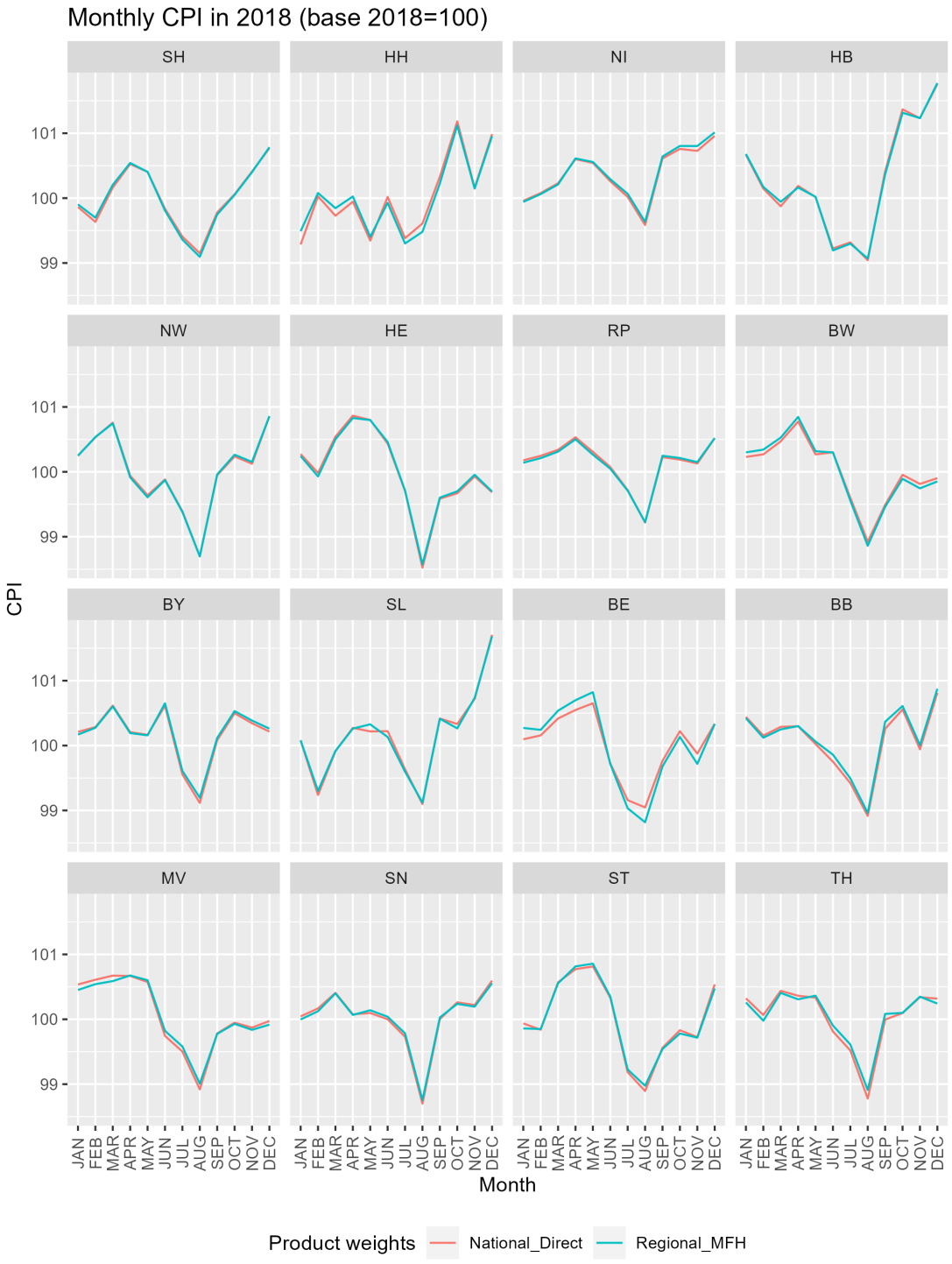


Figure B1: Estimation of regional CPI

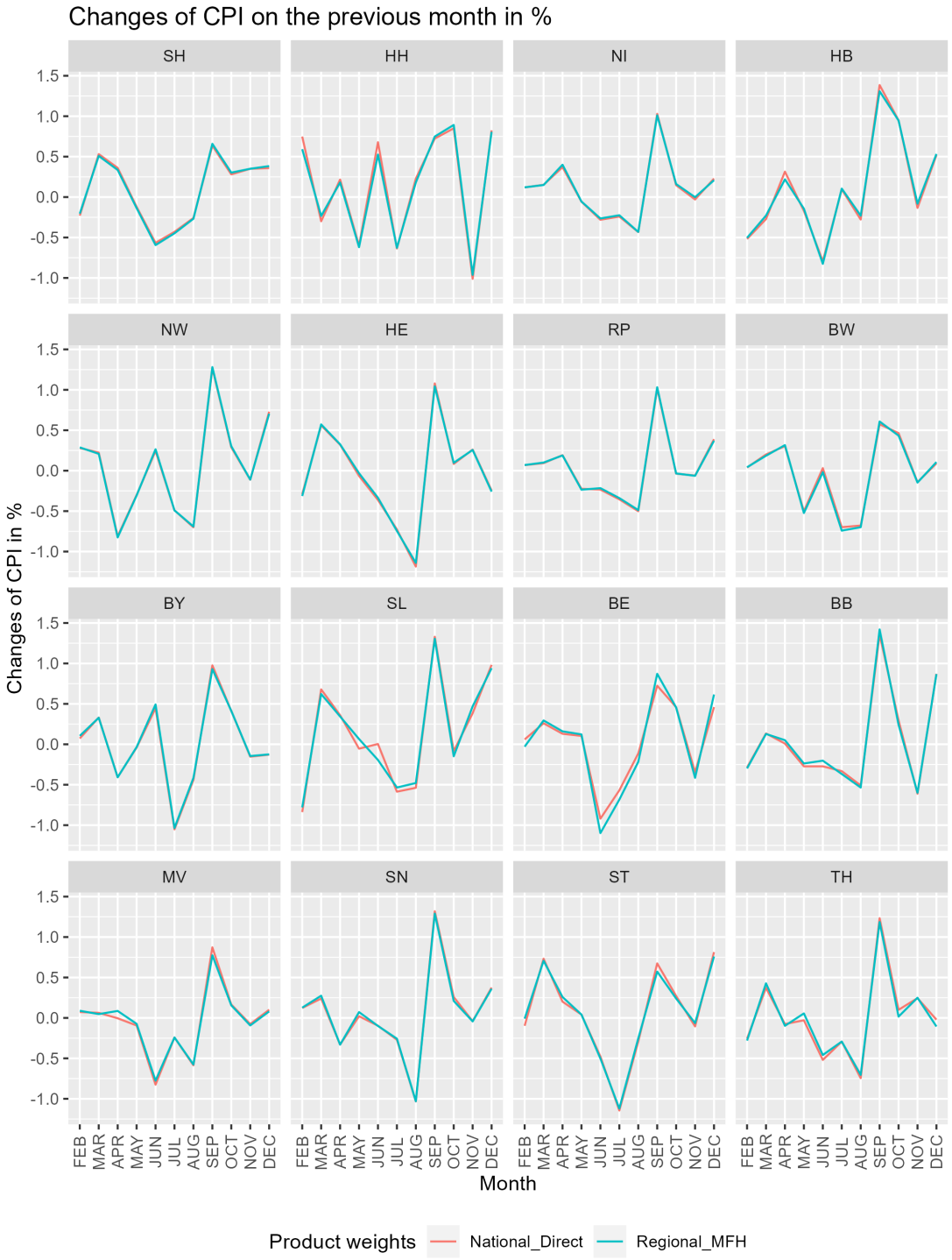


Figure B2: Changes of regional CPI from the previous month in %

Table B1: Description of state codes

Code	State	Code	State
SH	Schleswig-Holstein	HH	Hamburg
NI	Lower Saxony	HB	Bremen
NW	North Rhine-Westphalia	HE	Hesse
RP	Rhineland-Palatinate	BW	Baden-Württemberg
BY	Bavaria	SL	Saarland
BE	Berlin	BB	Brandenburg
MV	Mecklenburg-Vorpommern	SN	Saxony
ST	Saxony-Anhalt	TH	Thuringia

Table B2: Description of COICOP 4 codes

Code	Description	Code	Description
0111	Bread and cereals	0112	Meats and meat products
0113	Fish, fish products, and seafood	0114	Dairy products and eggs
0115	Oils and fats	0116	Fruits
0117	Vegetables	0118	Sugar, jam, marmalade, honey, syrup, chocolate, and sweets
0119	Other food products n.e.c.	0121	Coffee, tea, and cocoa drinks
0122	Mineral water, lemonades, fruit and vegetable juices	0211	Liquors
0212	Wine	0213	Beer
0220	Tobacco		

Chapter 3

Small area estimation using geospatial data based on transformed two-fold nested error regression models

Abstract

Fighting poverty starts with identifying where exactly poverty is the most severe by estimating poverty indicators. For a precise estimation of indicators at disaggregated regional levels, a small area estimation (SAE) approach is essential. SAE methods require a survey and a auxiliary dataset. By combining two datasets, SAE methods overcome the problem of small sample sizes and produce reliable estimates at a small area level. A census dataset is commonly used as auxiliary data to estimate poverty indicators. However, data protection laws often impede access and even a recent census may already be outdated because in developing countries, changes are as quick as they are dynamic. Therefore, an older census is inappropriate as auxiliary data. Geospatial data is a viable alternative since it is freely available, up to date, and covers all inhabited areas. Still, a central challenge remains: geospatial data is collected at grid level which is usually larger than a household but smaller than a small area. Since traditional SAE models use either a household level model or an area level model, we need to investigate which SAE model optimizes the advantages of grid level data. We suggest the two-fold nested regression model as it allows two random effects at different regional levels. More random effects capture the hierarchical data structure more accurately than standard SAE models. Additionally, we introduce transformations to the two-fold model to correct the violation in distributional model assumptions and to estimate ratio type indicators. We furthermore propose an estimation method for mean squared errors (MSE) with the transformed two-fold model. We show an efficiency gain of the two-fold model compared to the standard Fay-Herriot model, and the proposed MSE estimator is reasonable. Lastly, we apply the proposed transformed two-fold model to the geospatial data of Mozambique to estimate poverty indicators for 161 districts.

Keywords: head count ratio, Mozambique, satellite imagery, remote sensing data, subarea level model

3.1 Introduction

Fighting poverty presents considerable challenges. In fact, to formulate effective policy up to the regional level, decision makers need as much information as they can gather about where poverty is located. Accordingly, countries and organizations fighting poverty frequently estimate poverty indicators at a granular regional level. For an exact estimation of indicators at this disaggregated small area level, a

small area estimation (SAE) method is necessary. The reason for this is that at a small area level, a direct estimation of indicators from a survey with, for example, a Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) or a generalized regression estimator cannot produce reliable estimates due to unobserved areas or small sample sizes. SAE methods, on the other hand, combine survey with auxiliary data based on a model to produce more accurate estimations at small area levels (Rao and Molina, 2015; Morales et al., 2021). Commonly, SAE methods use a household income and expenditure survey and a census dataset as auxiliary data for the estimation of poverty indicators. It is, however, difficult to obtain a census dataset which accurately represents the current population: data protection policy may hinder access for researchers, and census data may already be outdated because a census is only carried out every ten years in most countries. Especially in developing countries facing dynamic structural changes and internal conflicts, an older census cannot accurately represent the current population. For example, the last official census of Lebanon was conducted in 1932 (Office of International Religious Freedom, 2022).

To overcome this problem, the use of geospatial data is an interesting viable alternative to census data. Geospatial data is publicly available, up to date, and covers all regions. Consequently, it has huge potential as auxiliary covariates in SAE models. Geospatial data is extracted from satellite imagery which is divided into multiple cells called grids. From the imagery, information about locations such as land cover, nighttime light, or rainfall is recorded for each grid. Current research shows empirical evidence that geospatial data is highly related to economic welfare and poverty (Steele et al., 2017; Yeh et al., 2020; Newhouse, 2023). Researchers also investigate the use of geospatial data in the SAE context. For example, while standard SAE models are defined either at household level or at the area level of interest, Newhouse et al. (2022) and Masaki et al. (2022) suggest to use a unit-context small area model which models the households' welfare variable using grid level geospatial covariates. Masaki et al. (2022) criticize the area level model because the variation in geospatial covariates is neglected through the aggregation of variables from grids to areas. However, Corral et al. (2021) show that the unit-context model has a bias that cannot be ignored. As a result, it is still unclear on which level we should define the model when we want to use grid level geospatial data.

We suggest a subarea level model as reasonable SAE model for geospatial data. Erciulescu et al. (2019) propose and investigate a hierarchical Bayesian subarea level model with geospatial auxiliary data to estimate harvested acreage at a small area level. The authors argue that the model at subarea level captures the hierarchical regional structure by defining additional random effects at the subarea level. Torabi and Rao (2014) extend the Fay-Herriot model (Fay and Herriot, 1979), which is one of the most famous area level models, and introduce the subarea level model. This subarea model is also known as two-fold nested regression model (Morales et al., 2021). The two-fold model can overcome the shortcomings of the unit-context model and area level models. In the present study, we extend the two-fold model with transformations to improve the performance of the estimation and estimate ratio type indicators. Moreover, we investigate the transformed two-fold model compared to the Fay-Herriot model. We also estimate poverty indicators at district level in Mozambique to show the performance of our model with real data.

Mozambique is one of the poorest countries in the world. After the end of the country's civil war in 1992, Mozambique achieved tentative stability both politically and economically. In the 2010s, however, Mozambique struggled again due to a resurgence of political violence, an insurgence by militant islamist forces, and an economic crisis caused by hidden debts in 2016. The outcome was a reversal of the successes made in the fight against poverty in the 2000s. It is obvious that Mozambique's achievements in fighting poverty are more volatile and inconsistent. Mozambique is changing at such a rapid pace, that the use of outdated auxiliary data will produce strong inaccuracies. Therefore, we use current geospatial auxiliary data to produce accurate estimates of poverty indicators at district level.

We proceed by introducing geospatial data and explaining the regional structure of our data in

Section 3.2. In Section 3.3, we propose a transformed two-fold nested error regression model and the mean squared error (MSE) estimation method for our approach in Section 3.4. In Section 3.5, we evaluate the newly introduced methods based on a model-based simulation study and apply our model to real data from Mozambique in Section 3.6.

3.2 Geospatial data and hierarchical data structure

The goal of this study is a precise estimation of poverty indicators at district level in Mozambique using geospatial auxiliary data. Specifically, we focus on the estimation of the mean household consumption and the head count ratio (HCR) (Foster et al., 1984). For the estimation, we use *Inquérito sobre orçamento familiar* from 2019/2020 (*IOF 2019/20* - household budget survey) performed by *Instituto Nacional de Estatistica* (*INE* - National Institute of Statistics) and funded by the World Bank as sample data. *IOF 2019/20* contains a variable measuring average household consumption per capita per day in Mozambican metical (MZN) that we use as the welfare variable for calculating the indicators of interest.

As auxiliary data, we use geospatial data in 2019 from Friedl and Sulla-Menashe (2022), Huffman et al. (2019), and Román et al. (2018). Geospatial data includes multiple geospatial variables which can be used as explanatory variables in the two-fold model, such as proportion of land covered by forests, proportion of croplands, maximum of nighttime light, or median of rainfall. These variables are extracted from satellite imagery. From raw satellite imagery, different types of maps are created such as land cover classification maps or maps of nighttime light. For example, Figure 3.1 describes how geospatial variables related to land cover information are collected for the province Maputo in Mozambique. Land cover variables are extracted from the land cover classification map. The land cover classification map is divided into multiple cells called grids and information like the proportion of lands covered by grass for each grid is recorded. Satellite imagery, as well as different maps extracted from imagery, are freely available, up to date, and cover the entire region of interest. Therefore, it is a fitting data source which can replace the census data after examining how it should be used in a SAE context.

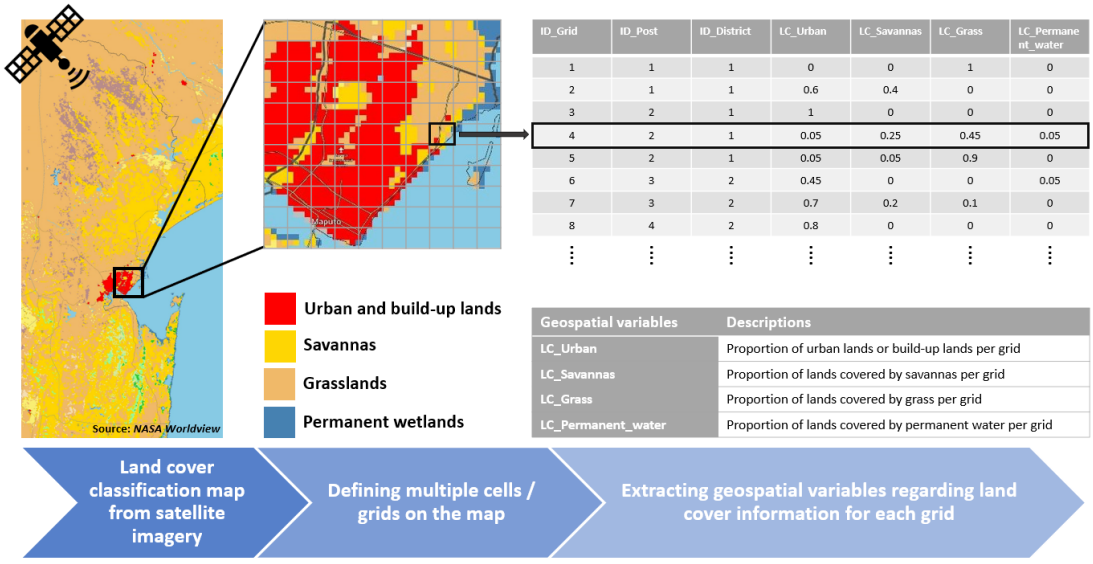


Figure 3.1: Collection of geospatial variables with an example of land cover variables in province Maputo of Mozambique (Source: *NASA Worldview*)

To find an efficient way to use geospatial data as alternative auxiliary data, we must understand the new data structure together with the geospatial data. With standard SAE methods, we use a sample dataset and an auxiliary census dataset collected at household level to obtain estimates of indicators

at area level. This means that there are two regional levels in the population for the SAE problem: households and areas. We want to use geospatial data collected at grid level and combine it with the household level survey data to estimate indicators at the area level. A single grid is commonly bigger than a household but smaller than an area, which means that we have a multilevel structure in our data. Between grids and our target areas, there are other possible subdivisions as shown in the gray colored level in Figure 3.2. When we have only grids as additional level between the target areas and the households, the data has a 3 level structure which corresponds to Figure 3.2 without the gray level. Therefore, we need to find a model which optimizes this multilevel hierarchical data structure to obtain precise estimation.

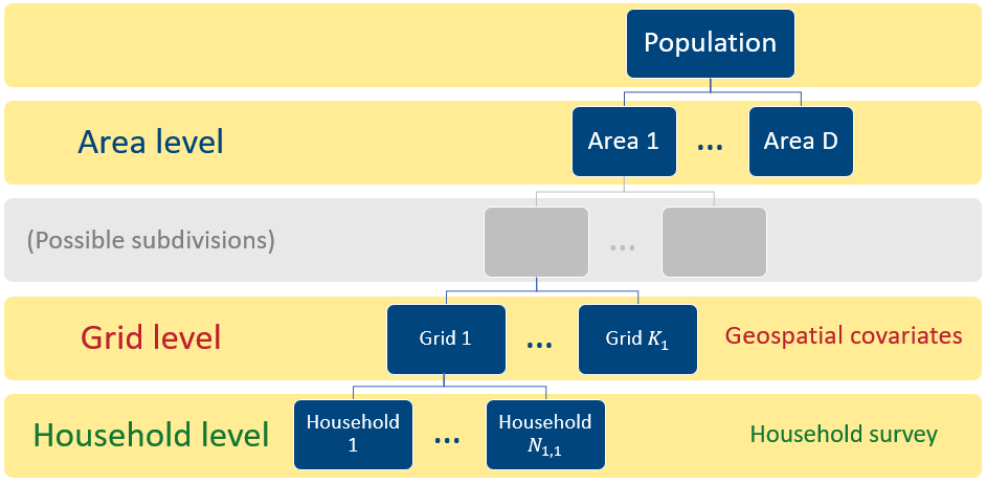


Figure 3.2: Hierarchical regional structure of the population with household survey and geospatial data in a SAE context

SAE methods use either a unit level model or an area level model. Since geospatial data does not contain household level information, we cannot use a unit level model. A unit-context model is also inadequate because it produces biased estimation results (Corral et al., 2021). We can try to fit a model at the area level using, for example, the Fay-Herriot model. To fit this model, we calculate the direct estimator of the target indicator from the survey data at area level and aggregate the grid level auxiliary variables to the area level from the grid level geospatial data. However, there are problems. First, the Fay-Herriot model includes random effects only at area level, which means that the specific effects at other levels, such as at grid level, are ignored. Furthermore, we lose information through the data aggregation from grids to areas (Masaki et al., 2022; Erciulescu et al., 2019).

We suggest to solve this problem by using a two-fold nested error regression model (Torabi and Rao, 2014) with transformations. This two-fold model is defined at subarea level which can be grid or any other proper subdivision between grids and areas. This model defines random effects not only at area level but also at subarea level. As a result, we fit the model at subarea level and take into account more regional specific effects by including two random effects. By using the auxiliary data at grid or subarea level, we reduce the loss of information in the geospatial auxiliary data. Moreover, the two-fold model with transformations helps to fulfill distributional assumptions and to ensure the interval of predicted values for indicators having a limited interval, such as proportions.

3.3 EBLUP under transformed two-fold nested error regression model

In this section, we briefly describe the two-fold nested error regression model by Torabi and Rao (2014) and propose our transformed two-fold model. Since we aim to estimate mean consumption which usually does not fulfill the normality assumptions and the HCR needs to be between 0 and 1, we explain the details of transformed two-fold models with a focus on the logarithmic and the arcsine transformation. To get the empirical best linear unbiased predictor (EBLUP) in the original scale, we need to back-transform the estimates. A naive back-transformation, however, creates a bias (Sugasawa and Kubokawa, 2017; Slud and Maiti, 2006). Therefore, we introduce the bias corrected back-transformation by Sugasawa and Kubokawa (2017) at the end of this section.

3.3.1 Transformed two-fold nested error regression model

Before we define the transformed two-fold model, we briefly introduce the two-fold model without transformation by Torabi and Rao (2014). Assume U is a finite population divided into multiple areas $i = 1, \cdots, D$. Each area i has multiple subareas $k = 1, \cdots, M_i$. In each subarea, there are N_{ik} households and the population size of area i is defined as $N_i = \sum_{k=1}^{M_i} N_{ik}$. The total population size is given by $N = \sum_{i=1}^{D} N_i$. Torabi and Rao (2014) define the indicator of interest θ in area i in subarea k (θ_{ik}) in a two-fold nested error regression model as follows

$$\theta_{ik} = x_{ik}^T\beta + v_i + u_{ik}, \text{ for } i = 1, \cdots, D \text{ and } k = 1, \cdots, M_i$$

with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $u_{ik} \overset{iid}{\sim} N(0, \sigma_u^2)$. x_{ik} is a $p \times 1$ vector of independent variables of subarea k in area i and β is a $p \times 1$ vector of fixed effects. θ_{ik} includes two random effects: at area level (v_i) and at subarea level (u_{ik}). v_i and u_{ik} are independent of each other.

The direct estimator of θ_{ik} can be calculated based on a survey dataset. In the survey, we have d different areas and m_i different subareas within an area i. There are n_{ik} units in subarea k in area i. The sample size of area i is given by $n_i = \sum_{k=1}^{m_i} n_{ik}$ and the total sample size is $n = \sum_{i=1}^{d} n_i$. From the sample dataset, the appropriate direct estimator of the indicator ($\widehat{\theta}_{ik}^{Dir}$) is calculated, such as the HT estimator. The direct estimator $\widehat{\theta}_{ik}^{Dir}$ is defined as the sum of the true indicator and the sampling error as follows

$$\widehat{\theta}_{ik}^{Dir} = \theta_{ik} + e_{ik} = x_{ik}^T\beta + v_i + u_{ik} + e_{ik}, \text{ for } i = 1, \cdots, d \text{ and } k = 1, \cdots, m_i$$

with known sampling errors $e_{ik} \overset{ind}{\sim} N(0, \sigma_{e_{ik}}^2)$.

Now, we define a transformed two-fold model

$$\widehat{\theta}_{ik}^{Dir,*} = T(\widehat{\theta}_{ik}^{Dir}) = \theta_{ik}^* + e_{ik}^* = x_{ik}^T\beta + v_i + u_{ik} + e_{ik}^*, \tag{3.1}$$

with $e_{ik}^* \sim N(0, V(\widehat{\theta}_{ik}^{Dir,*}))$ where $\widehat{\theta}_{ik}^{Dir,*}$ denotes transformed direct estimates and $T(\cdot)$ is a transformation function. Since the model is transformed, $x_{ik}^T\beta + v_i + u_{ik}$ is no longer equal to the true indicator θ_{ik}. Therefore, we define $x_{ik}^T\beta + v_i + u_{ik}$ as θ_{ik}^* in a transformed two-fold model and the true indicator θ_{ik} is defined as $\theta_{ik} = T^{-1}(\theta_{ik}^*)$. Our goal is to find the EBLUP of θ_{ik} based on this transformed model. Additionally, e_{ik}^* is also no longer equal to the sampling error of the direct estimates. This means that we must calculate the variance of e_{ik}^* because the two-fold model assumes the known variance of e_{ik}^*. We can extract these variances of $\widehat{\theta}_{ik}^{Dir,*}$ from the known sampling variances of $\widehat{\theta}_{ik}^{Dir}$, $\sigma_{e_{ik}}^2$ by using

Taylor approximation

$$V(\widehat{\theta}_{ik}^{Dir,*}) = V(T(\widehat{\theta}_{ik}^{Dir})) \approx \left(T'(E(\widehat{\theta}_{ik}^{Dir}))\right)^2 \cdot V(\widehat{\theta}_{ik}^{Dir}) \tag{3.2}$$

$$\approx \left(T'(E(\widehat{\theta}_{ik}^{Dir}))\right)^2 \cdot \sigma_{e_{ik}}^2 \tag{3.3}$$

where $T'(\cdot)$ is the first order derivative of the transformation function (Neves et al., 2013; Rao and Molina, 2015). For the transformation functions that we use, $V(\widehat{\theta}_{ik}^{Dir,*})$ can be simplified as

$$V(\widehat{\theta}_{ik}^{Dir,*}) = \begin{cases} (\widehat{\theta}_{ik}^{Dir})^{-2} \cdot \sigma_{e_{ik}}^2 & \text{, if } T(\widehat{\theta}_{ik}^{Dir}) = \log(\widehat{\theta}_{ik}^{Dir}) \\ 1/(4\widetilde{n}_{ik}) & \text{, if } T(\widehat{\theta}_{ik}^{Dir}) = \sin^{-1}\sqrt{\widehat{\theta}_{ik}^{Dir}}, \end{cases} \tag{3.4}$$

where $\widetilde{n}_{ik}$ is the effective sample size for subarea k in area i. For details of the simplification, see Neves et al. (2013) and Jiang et al. (2001). These approximated sampling variances are the new known variances of $e_{ik}^* \sim N(0, V(\widehat{\theta}_{ik}^{Dir,*}))$ in the transformed two-fold model.

The two-fold model defined in Equation (3.1) can be expressed at area level for all areas $i = 1, \cdots, d$ using matrix notation as follows

$$\begin{aligned} \widehat{\theta}_i^{Dir,*} =& X_i\beta + v_i 1_{m_i} + u_i + e_i^*, \\ =& X_i\beta + Z_i b_i + e_i^* \end{aligned} \tag{3.5}$$

where $\widehat{\theta}_i^{Dir,*} = (\widehat{\theta}_{i1}^{Dir,*}, \cdots, \widehat{\theta}_{im_i}^{Dir,*})^T$, $u_i = (u_{i1}, \cdots, u_{im_i})^T$ and $e_i^* = (e_{i1}^*, \cdots, e_{im_i}^*)^T$. X_i denotes a $m_i \times p$ design matrix with $X_i = (x_{ik}^T, \cdots, x_{im_i}^T)^T$. Z_i is $(1_{m_i}|I_{m_i})$ with 1_{m_i} as the $m_i \times 1$ vector of ones and I_{m_i} as the m_i dimensional identity matrix and $b_i = (v_i, u_i^T)^T$. Then, the covariance matrix of the model in Equation (3.5) is

$$V_i = R_i + Z_i G_i Z_i^T$$

where $R_i = diag(V(\widehat{\theta}_{i1}^{Dir,*}), \cdots, V(\widehat{\theta}_{im_i}^{Dir,*}))$ and

$$G_i = \begin{pmatrix} \sigma_v^2 & 0 \\ 0 & \sigma_u^2 I_{m_i} \end{pmatrix}.$$

3.3.2 Estimation of the EBLUP at subarea level

To obtain the EBLUP of θ_{ik}^*, the model parameters must be defined and estimated. In the case of known σ_u^2 and σ_v^2, the best linear unbiased estimator (BLUE) of β ($\widetilde{\beta}$) and the BLUP of v_i and u_{ik} ($\widetilde{v}_i$, $\widetilde{u}_{ik}$) are given by

$$\widetilde{\beta} = \left(\sum_{i=1}^{d} X_i^T V_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{d} X_i^T V_i^{-1} \widehat{\theta}_i^{Dir,*}\right), \tag{3.6}$$

$$\widetilde{v}_i = \gamma_i \left(\frac{1}{\sum_{k=1}^{m_i}\gamma_{ik}} \sum_{k=1}^{m_i} \gamma_{ik}\widehat{\theta}_{ik}^{Dir,*} - \frac{1}{\sum_{k=1}^{m_i}\gamma_{ik}} \sum_{k=1}^{m_i} \gamma_{ik} x_{ik}\widehat{\beta}\right), \tag{3.7}$$

$$\widetilde{u}_{ik} = \gamma_{ik}(\widehat{\theta}_{ik}^{Dir,*} - x_{ik}^T\hat{\beta}) - \gamma_i\gamma_{ik}\left(\frac{1}{\sum_{k=1}^{m_i}\gamma_{ik}} \sum_{k=1}^{m_i} \gamma_{ik}\widehat{\theta}_{ik}^{Dir,*} - \frac{1}{\sum_{k=1}^{m_i}\gamma_{ik}} \sum_{k=1}^{m_i} \gamma_{ik} x_{ik}\widehat{\beta}\right), \tag{3.8}$$

with

$$\gamma_{ik} = \frac{\sigma_u^2}{\sigma_u^2 + V(\widehat{\theta}_{ik}^{Dir,*})} \text{ and } \gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + (\sigma_u^2 / \sum_{k=1}^{m_i} \gamma_{ik})}.$$

Let $\widetilde{\theta}_{ik}^*$ the BLUP of θ_{ik}^*. Then, $\widetilde{\theta}_{ik}^*$ is obtained as follows

$$\widetilde{\theta}_{ik}^* = \begin{cases} x_{ik}^T \widetilde{\beta} + \widetilde{v}_i + \widetilde{u}_{ik}, & \text{for in sample subareas,} \\ x_{ik}^T \widetilde{\beta} + \widetilde{v}_i, & \text{for out-of-sample subareas.} \end{cases}$$

The EBLUP of θ_{ik}^* at subarea level is denoted by $\widehat{\theta}_{ik}^{TF,*}$ and is obtained by replacing parameters with its estimators. First, σ_u and σ_v are replaced by $\hat{\sigma}_u$ and $\hat{\sigma}_v$ to obtain $\widehat{\gamma}_{ik}$ and $\widehat{\gamma}_i$

$$\widehat{\gamma}_{ik} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + V(\widehat{\theta}_{ik}^{Dir,*})} \text{ and } \widehat{\gamma}_i = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + (\hat{\sigma}_u^2 / \sum_{k=1}^{m_i} \widehat{\gamma}_{ik})}. \tag{3.9}$$

$\hat{\sigma}_u$ and $\hat{\sigma}_v$ can be estimated with different methods such as weighted least squares, maximum likelihood, or residual maximum likelihood. $\widehat{\widetilde{\beta}}$ is obtained by inserting these estimated variance components into the variance matrix (V_i) in Equation (3.6). We calculate $\widehat{\widetilde{v}}_i$ and $\widehat{\widetilde{u}}_{ik}$ by replacing γ_{ik} and γ_i in Equation (3.7) - (3.8) with $\widehat{\gamma}_{ik}$ and $\widehat{\gamma}_i$. Then, $\widehat{\theta}_{ik}^{TF,*}$ is given by

$$\widehat{\theta}_{ik}^{TF,*} = \begin{cases} x_{ik}^T \widehat{\widetilde{\beta}} + \widehat{\widetilde{v}}_i + \widehat{\widetilde{u}}_{ik}, & \text{for in sample subareas,} \\ x_{ik}^T \widehat{\widetilde{\beta}} + \widehat{\widetilde{v}}_i, & \text{for out-of-sample subareas.} \end{cases} \tag{3.10}$$

3.3.3 Bias corrected back-transformation

Since $\widehat{\theta}_{ik}^{TF,*}$ is the EBLUP of θ_{ik}^* from the transformed model in Equation (3.1), we need to back-transform $\widehat{\theta}_{ik}^{TF,*}$ to obtain the EBLUP in the original scale ($\widehat{\theta}_{ik}^{TF}$). By using the inverse of the transformation function, we can estimate the naive back-transformed EBLUP at subarea level and area level as follows

$$\begin{aligned} \widehat{\theta}_{ik}^{TF,\ Naive} &= T^{-1}(\widehat{\theta}_{ik}^{TF,*}), \\ \widehat{\theta}_{i}^{TF,\ Naive} &= \frac{1}{N_i} \sum_{k=1}^{M_i} N_{ik} \widehat{\theta}_{ik}^{TF,\ Naive}. \end{aligned} \tag{3.11}$$

However, the simple naive back-transformation causes bias (Sugasawa and Kubokawa, 2017; Slud and Maiti, 2006). For the bias correction, Sugasawa and Kubokawa (2017) propose a bias corrected back-transformation using the conditional expectation $\widetilde{\theta}_{ik} = E(T^{-1}(\theta_{ik}^*)|\widehat{\theta}_{ik}^{Dir})$ where $\widetilde{\theta}_{ik}$ is the BLUP of θ_{ik}. The BLUP of θ_{ik}^* given $\widehat{\theta}_{ik}^{Dir}$ is defined in Equation (3.10) and has the following distribution

$$\widetilde{\theta}_{ik}^* \sim N(\widetilde{\theta}_{ik}^*, g_{ik})$$

with

$$g_{ik} = \begin{cases} (1-\gamma_{ik})^2[\sigma_u^2 + (1-\gamma_i)\sigma_v^2] + \gamma_{ik}^2 \cdot V(\widehat{\theta}_{ik}^{Dir,*}) & \text{for in sample subarea } k \\ \sigma_u^2 \left[1 + \frac{\gamma_i}{\sum_{k=1}^{m_i} \gamma_{ik}}(1 - 2\gamma_{ik})\right] & \text{for out-of-sample subarea } k. \end{cases} \tag{3.12}$$

For the detailed derivation of g_{ik}, see Torabi and Rao (2014). Based on this distribution, the $\tilde{\theta}_{ik}$ is obtained as follows

$$\widetilde{\theta}_{ik} = \int_{-\infty}^{\infty} T^{-1}(t)\phi(t; \widetilde{\theta}^*_{ik},\ g_{ik})dt, \tag{3.13}$$

where $\phi(\cdot; \widetilde{\theta}^*_{ik},\ g_{ik})$ is the density function of the normal distribution $N(\widetilde{\theta}^*_{ik},\ g_{ik})$. By replacing $(\widetilde{\theta}^*_{ik}, g_{ik})$ with $(\widehat{\theta}^{TF,*}_{ik}, \widehat{g}_{ik})$, in Equation (3.13) the bias corrected back-transformed EBLUP is obtained by

$$\widehat{\theta}^{TF}_{ik} = \int_{-\infty}^{\infty} T^{-1}(t)\phi(t; \widehat{\theta}^{TF,*}_{ik},\ \widehat{g}_{ik})dt. \tag{3.14}$$

$\widehat{g}_{ik}$ is obtained by replacing $(\gamma_{ik}, \gamma_i, \sigma_u, \sigma_v)$ in Equation (3.12) with $(\widehat{\gamma}_{ik}, \widehat{\gamma}_i, \hat{\sigma}_u, \hat{\sigma}_v)$. $\widehat{\theta}^{TF}_{ik}$ can be expressed with the transformations we used as follows

$$\widehat{\theta}^{TF}_{ik} = \begin{cases} \exp\left[\widehat{\theta}^{TF,*}_{ik} + 0.5\widehat{g}_{ik}\right], & \text{for } T(\widehat{\theta}^{Dir}_{ik}) = \log(\widehat{\theta}^{Dir}_{ik}) \\ \int_{-\infty}^{\infty} \sin^2(t) \frac{1}{2\pi\hat{g}_{ik}} \exp\left[-(t - \widehat{\theta}^{TF,*}_{ik})^2/(2\widehat{g}_{ik})\right] dt, & \text{for } T(\widehat{\theta}^{Dir}_{ik}) = \sin^{-1}\sqrt{\widehat{\theta}^{Dir}_{ik}}. \end{cases}$$

Based on the bias corrected EBLUP at subarea level, we estimate the bias corrected EBLUP at area level as

$$\widehat{\theta}^{TF}_{i} = \frac{1}{N_i}\sum_{k=1}^{M_i} N_{ik}\widehat{\theta}^{TF}_{ik}.$$

3.4 MSE estimation

To measure the uncertainty of the EBLUP $\widehat{\theta}^{TF}_{ik}$ at subarea and $\widehat{\theta}^{TF}_{i}$ at area level, we propose an MSE estimation with the parametric bootstrap method based on González-Manteiga et al. (2008) and Hadam et al. (2023). Using a bootstrap method to estimate MSE is a reasonable choice since the derivation of the analytical MSE for general transformations is challenging because of its complexity. The parametric bootstrap procedure to estimate the MSE is stepwise explained as follows

1. Transform $\widehat{\theta}^{Dir}_{ik}$ to obtain $\widehat{\theta}^{Dir,\ *}_{ik}$. Calculate the variance of e^*_{ik} (i.e., $V(\widehat{\theta}^{Dir,*}_{ik})$) using Equation (3.3) or Equation (3.4).
2. Fit the model in Equation (3.1) with the obtained $\widehat{\theta}^{Dir,*}_{ik}$ and $V(\widehat{\theta}^{Dir,*}_{ik})$ to estimate the model parameters $\widehat{\widetilde{\beta}}, \hat{\sigma}_v$, and $\hat{\sigma}_u$.
3. For $b = 1, \cdots, B$:
   (a) Generate pseudo random effects at area level ($v^{(b)}_i \sim N(0, \hat{\sigma}^2_v)$) and subarea level ($u^{(b)}_{ik} \sim N(0, \hat{\sigma}^2_u)$), and pseudo error ($e^{*,(b)}_{ik} \sim N(0, V(\widehat{\theta}^{Dir,*}_{ik}))$) using the obtained $\hat{\sigma}_v, \hat{\sigma}_u$, and $V(\widehat{\theta}^{Dir,*}_{ik})$ from step 1-2.
   (b) Obtain the pseudo true indicator in the transformed scale ($\theta^{*,(b)}_{ik}$) and the transformed direct estimate of the indicator ($\widehat{\theta}^{Dir,*,(b)}_{ik}$) for the pseudo population as follows

$$\begin{aligned} \theta^{*,(b)}_{ik} &= x_{ik}\widehat{\widetilde{\beta}} + v^{(b)}_i + u^{(b)}_{ik} \\ \widehat{\theta}^{Dir,*,(b)}_{ik} &= x_{ik}\widehat{\widetilde{\beta}} + v^{(b)}_i + u^{(b)}_{ik} + e^{*,(b)}_{ik}. \end{aligned}$$

   (c) Transform the $V(\widehat{\theta}^{Dir,*}_{ik})$ using the inverse function of the Taylor approximation in Equation (3.3) to get the adjusted bootstrap variances of the direct indicators in the original scale

as follows

$$V(\widehat{\theta}_{ik}^{Dir})^{(b)} = \left(T'(E(\widehat{\theta}_{ik}^{Dir}))\right)^{-2} \cdot V(\widehat{\theta}_{ik}^{Dir,*}).$$

(d) Obtain the pseudo true indicator in the original scale using the inverse transformation function at subarea level

$$\theta_{ik}^{(b)} = T^{-1}(x_{ik}\widehat{\beta} + v_i^{(b)} + u_{ik}^{(b)}).$$

(e) Find the pseudo true indicator at area level based on $\theta_{ik}^{(b)}$ values as follows

$$\theta_i^{(b)} = \frac{1}{N_i}\sum_{k=1}^{M_i} N_{ik}\theta_{ik}^{(b)}.$$

(f) Fit the model in Equation (3.1) with $\widehat{\theta}_{ik}^{Dir,*,(b)}$ and $V(\widehat{\theta}_{ik}^{Dir})^{(b)}$ to find the two-fold EBLUP in the original scale at subarea level $\widehat{\theta}_{ik}^{TF,(b)}$ and at area level $\widehat{\theta}_i^{TF,(b)}$ as described in Section 3.3.

4. The MSE estimator is calculated as follows

$$\widehat{\text{MSE}}(\widehat{\theta}_{ik}^{TF}) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_{ik}^{TF,(b)} - \theta_{ik}^{(b)}\right)^2 \text{ for subareas}$$

$$\widehat{\text{MSE}}(\widehat{\theta}_i^{TF}) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_i^{TF,(b)} - \theta_i^{(b)}\right)^2 \text{ for areas.}$$

The root mean squared error (RMSE) estimates are obtained by taking the square root of MSE estimates.

3.5 Model-based simulation study

For the evaluation of the proposed EBLUP under a transformed two-fold model, we conduct a model-based simulation study. The purpose of this simulation study is to investigate the performance of the EBLUP with the transformed two-fold model compared to the standard EBLUP with the transformed Fay-Herriot model applied to multilevel structured data. For the simulation data, we assume a three level data structure which includes a household level, a subarea level, and an area level. Additionally, we aim to show that the proposed MSE estimator of the two-fold EBLUP performs well.

3.5.1 Simulation setting

In the simulation study, we assume a finite population U divided into $D = 100$ areas. The number of subareas in the ith area is denoted by M_i which is equal to 10 for all areas $i = 1, ..., D$. $N_{ik} = 20$ units exist in all subareas $k = 1, \cdots, M_i$ of all areas. Therefore, there are in total 1000 subareas in the population, $N_i = \sum_{k=1}^{M_i} N_{ik} = 200$ units in each area, and $N = \sum_{i=1}^{D} N_i = 20000$ in the population. The population data is generated at unit level using the model defined in Table 3.1. A detailed distribution of the variables used in the data generating process (DGP) is also given in the same table. With the same DGP, $L = 500$ Monte Carlo population datasets are randomly generated. The generated y variable has a highly right skewed distribution similarly to welfare variables in real data.

In each population data, we assume that there are random effects not only at the area level, but also at the subarea level. From each generated population data, a sample dataset is randomly chosen. 25%

of subareas in the population dataset are not sampled. The sample size of subareas is denoted by n_{ik}. n_{ik} is between 2 and 20 for $i = 1, \cdots, d$ and $k = 1, \cdots, m_i$ where d denotes the number of sampled areas and m_i is the number of sampled subareas in ith area. The sample size of areas is defined as $n_i = \sum_{k=1}^{m_i} n_{ik}$ and is between 29 and 108. The sum of n_i gives the total sample size of $n = 7538$. Let y_{ikj} be a welfare variable such as income or consumption of the jth household in the kth subarea of the ith area. x_{ikj} is a vector including the characteristics of the corresponding household.

Table 3.1: DGP and distribution of variables used in the model-based simulation study

Model	x_{ikj}	μ_{ik}	v_i	u_{ik}	ε_{ikj}
$y_{ikj} = \exp(9 - x_{ikj} + v_i + u_{ik} + \varepsilon_{ikj})$	$N(\mu_{ik}, 0.5^2)$	$U(2,3)$	$N(0, 0.4^2)$	$N(0, 0.6^2)$	$N(0, 0.8^2)$

To examine the performance of the two-fold EBLUP, we estimate two indicators: the mean of y and the HCR at area level. The HCR is defined as follows

$$\text{HCR}_i = \frac{1}{N_i} \sum_{k=1}^{M_i} \sum_{j=1}^{N_{ik}} I(y_{ikj} \leq t),$$

where t is the poverty threshold. $I(\cdot)$ is an indicator which returns 1 if the condition $y_{ikj} \leq t$ is fulfilled, otherwise it returns 0. Moreover, we estimate the same indicators based on the Fay-Herriot model for a comparison with the two-fold EBLUP. Both models require direct estimates of the indicators and corresponding variances of the direct estimates. Therefore, we calculate the HT estimator of these indicators at the subarea level for the two-fold model and at the area level for the Fay-Herriot model. The threshold t is set at 60% of the median y in the sample data for the estimation of the HCR. $\widehat{\text{Mean}}_i^{Dir}$ and $\widehat{\text{Mean}}_{ik}^{Dir}$ denote the HT estimator of the mean at the area and at the subarea level accordingly. Then, the HT estimates of the HCR are denoted by $\widehat{\text{HCR}}_i^{Dir}$ at the area level and by $\widehat{\text{HCR}}_{ik}^{Dir}$ at the subarea level. Furthermore, the analytical variances of the HT estimators are also estimated at both levels for both estimated indicators.

The two-fold model as well as the Fay-Herriot model assume normal distributed errors. However, we generate the simulation data with the DGP in Table 3.1 so that it violates the normality assumptions. The target variables of the two-fold and Fay-Herriot model to estimate the mean are the direct estimates of y, (i.e., $\widehat{\text{Mean}}_{ik}^{Dir}$ and $\widehat{\text{Mean}}_i^{Dir}$) and they show that the normality assumptions may not be fulfilled. Their detailed skewness and kurtosis values are summarized in Table C1 in the Appendix and the distributions of $\widehat{\text{Mean}}_{ik}^{Dir}$ and $\widehat{\text{Mean}}_i^{Dir}$ in one simulation dataset are shown in Figure C1 in the Appendix. To correct the violation in the normality assumptions, we apply the logarithmic transformation to the models for the mean estimation. Furthermore, we apply the arcsine transformation to the square root of the HCR direct estimator before we fit the models for HCR so that the predicted values are between 0 and 1 (Carter and Rolph, 1974; Jiang et al., 2001; Raghunathan et al., 2007; Schmid et al., 2017; Hadam et al., 2023).

3.5.2 Simulation results

For the estimation of indicators, we calculate 3 different EBLUP: EBLUP based on the transformed two-fold model with the bias corrected back-transformation (*TF*), EBLUP based on the transformed two-fold model with the naive back-transformation (*TF_naive*), and EBLUP based on the transformed Fay-Herriot model (*FH*). For the *FH*, the bias corrected back-transformation is based on Slud and Maiti (2006) and Sugasawa and Kubokawa (2017). We use the analytical MSE by Slud and Maiti (2006) for the mean and the parametric bootstrap method for the HCR by Hadam et al. (2023) for *FH*. The MSE of the EBLUP with the two-fold models is estimated by the proposed parametric bootstrap method in

Section 3.3. The number of bootstraps is set at $B = 500$ for the parametric bootstrap method.

For the comparison of the estimated indicators, we use relative bias (RB) and RMSE as quality measurements defined by

$$\begin{aligned} \text{RB}_i &= \frac{1}{L}\sum_{l=1}^{L}\frac{\widehat{\theta}_i^{(l)} - \theta_i^{(l)}}{\theta_i^{(l)}} \cdot 100, \\ \text{RMSE}_i &= \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\theta}_i^{(l)} - \theta_i^{(l)}\right)^2}, \end{aligned} \tag{3.15}$$

where $\widehat{\theta}_i^{(l)}$ is either the two-fold EBLUP or the Fay-Herriot EBLUP of the ith area in the lth simulation dataset. For the mean indicator, the *FH* is the least biased among all three models as shown in Table 3.2. Still, the *TF* for the mean has also very small RB values between approximately -4% and 3%. Furthermore, *TF* has the smallest RMSE for the mean indicator according to Figure 3.3. The interesting point is that the effect of the bias corrected back-transformation is clearly noticeable in terms of both measures RB and RMSE. The HCR of the *TF* shows the smallest RB and RMSE values among all three estimates. In the HCR case, the difference of bias corrected back-transformation and naive back-transformation is not as large as in the mean indicator case but the improvement in estimation is still clear.

Table 3.2: Relative bias in % of EBLUP based on different models

Indicator	EBLUP	Min	1.Q	Median	Mean	3.Q	Max
Mean	*TF*	-4.472	-2.678	-1.992	-1.844	-1.126	2.942
	TF_naive	-13.541	-9.704	-8.548	-8.515	-7.089	-5.234
	FH	-1.581	-0.520	0.019	0.005	0.516	1.482
HCR	*TF*	-3.113	-1.307	-0.516	-0.344	0.211	5.190
	TF_naive	-7.009	-5.309	-4.572	-4.721	-4.184	-2.453
	FH	0.329	1.760	2.262	2.326	2.792	4.963

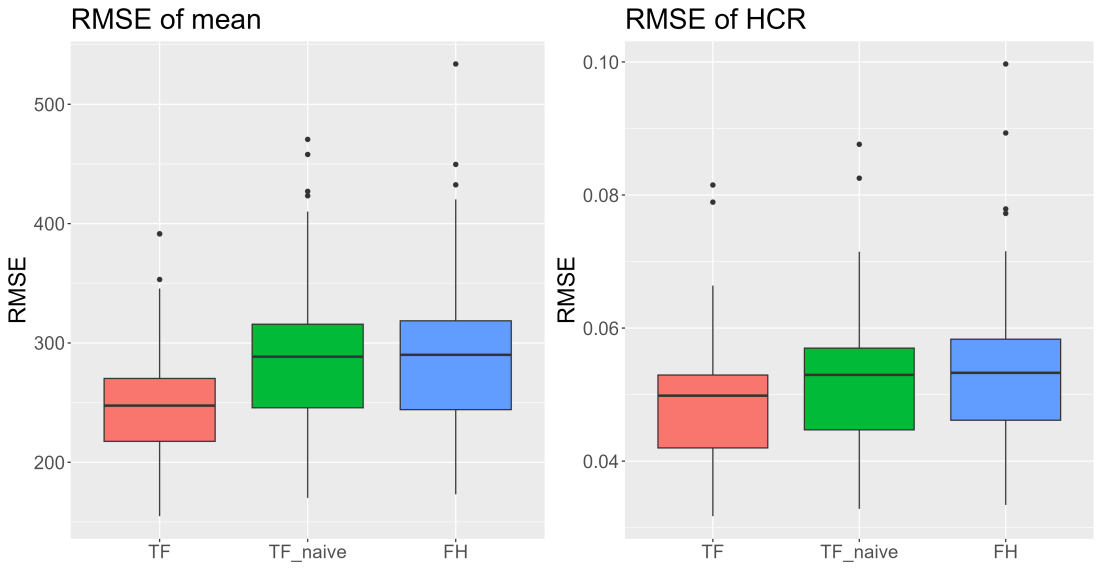


Figure 3.3: RMSE values of EBLUP based on different models

Next, we examine the performance of the proposed bootstrap MSE estimation in Section 3.4 for *TF*.

For the evaluation, we investigate the relative bias of RMSE (RB-RMSE) defined as

$$\text{RB-RMSE}_i = \frac{\sqrt{\frac{1}{L}\sum_{l=1}^{L}\widehat{\text{MSE}}(\widehat{\theta}_i^{(l)})} - \text{RMSE}_i}{\text{RMSE}_i} \cdot 100.$$

Note that RMSE_i is the empirical RMSE defined in Equation (3.15). Table 3.3 shows the summary of RB-RMSE values. For both indicators, the RB-RMSE values are low enough. Unlike MSE estimates of the HCR for *TF*, MSE estimates of the mean are generally smaller than the empirical RMSE values. In Figure 3.4, we can see on the upper line plot that the blue line is generally located slightly beneath the red line, while the lower line plot for HCR shows the opposite. Still, the blue line tracks the red line fairly well for both indicators. Based on these quality measures, we conclude that the proposed bootstrap MSE estimation is a reasonable method to estimate the MSE of *TF*.

Table 3.3: Relative bias in % of RMSE estimates

Indicator	Measurement	Min	1.Q	Median	Mean	3.Q	**Max**
Mean	RB-RMSE	-32.556	-11.243	-4.040	-5.078	0.594	20.255
HCR	RB-RMSE	-6.394	5.542	8.579	9.377	13.426	23.096

3.6 Application

In this section, we apply the proposed method in Sections 3.3 and 3.4 to Mozambique data for the estimation of poverty indicators at a disaggregated level. Mozambique is located in southeastern Africa. After the end of a ten-year long war of independence against former colonial power Portugal in 1974, the country fought a civil war that lasted until 1992. This left Mozambique one of the poorest countries in the world. However, the government of Mozambique still managed to enact a number of economic reforms beginning in 1987, which, together with continued domestic peace-mongering and resettlement of internally displaced persons, contributed to considerably high average real gross domestic product (GDP) growth rates. The average GDP growth rate amounts to 8% between 1996 and 2015 (IMF, 2023).

Mozambique's strong growth and positive economic development crashed in the 2010s due to several factors. First, there was a resurgence of political violence due to renewed anti-state activities by *Resistência Nacional Moçambicana* (*RENAMO* - Mozambican National Resistance), a political party and military group fighting the ruling *Frente de Libertação de Moçambique* (*FRELIMO* - Liberation Front of Mozambique) party. Second, there was another insurgence by militant Islamist forces in the natural resource rich northern province Cabo Delgado, which hindered the exploitation of natural resources. Third, in 2016, hidden debts of state-owned enterprises in Mozambique came to light, effectively destroying foreign investors' trust in Mozambique and drastically increasing the country's debt to GDP ratio (World Bank, 2016). Finally, Mozambique was also hit hard by the COVID pandemic, and especially rural regions have suffered immensely (World Bank, 2023b). When looking at GDP per capita in 2023 (IMF, 2023), Mozambique is still one of the poorest countries in the world despite its natural resources. In addition to general poverty, Mozambique faces numerous development challenges related to limited job creation, public education, health services, and public safety (World Bank, 2023b).

The first step to finding effective policies for solving challenges like the ones Mozambique deals with is to identify where exactly they are. In other words, one needs to know where exactly poverty is located. For the identification of problematic locations, estimation of poverty indicators at disaggregated regional levels is a useful tool and a natural starting point. Therefore, we aim to estimate the mean consumption per capita per day and the HCR at district level using an appropriate SAE method with the

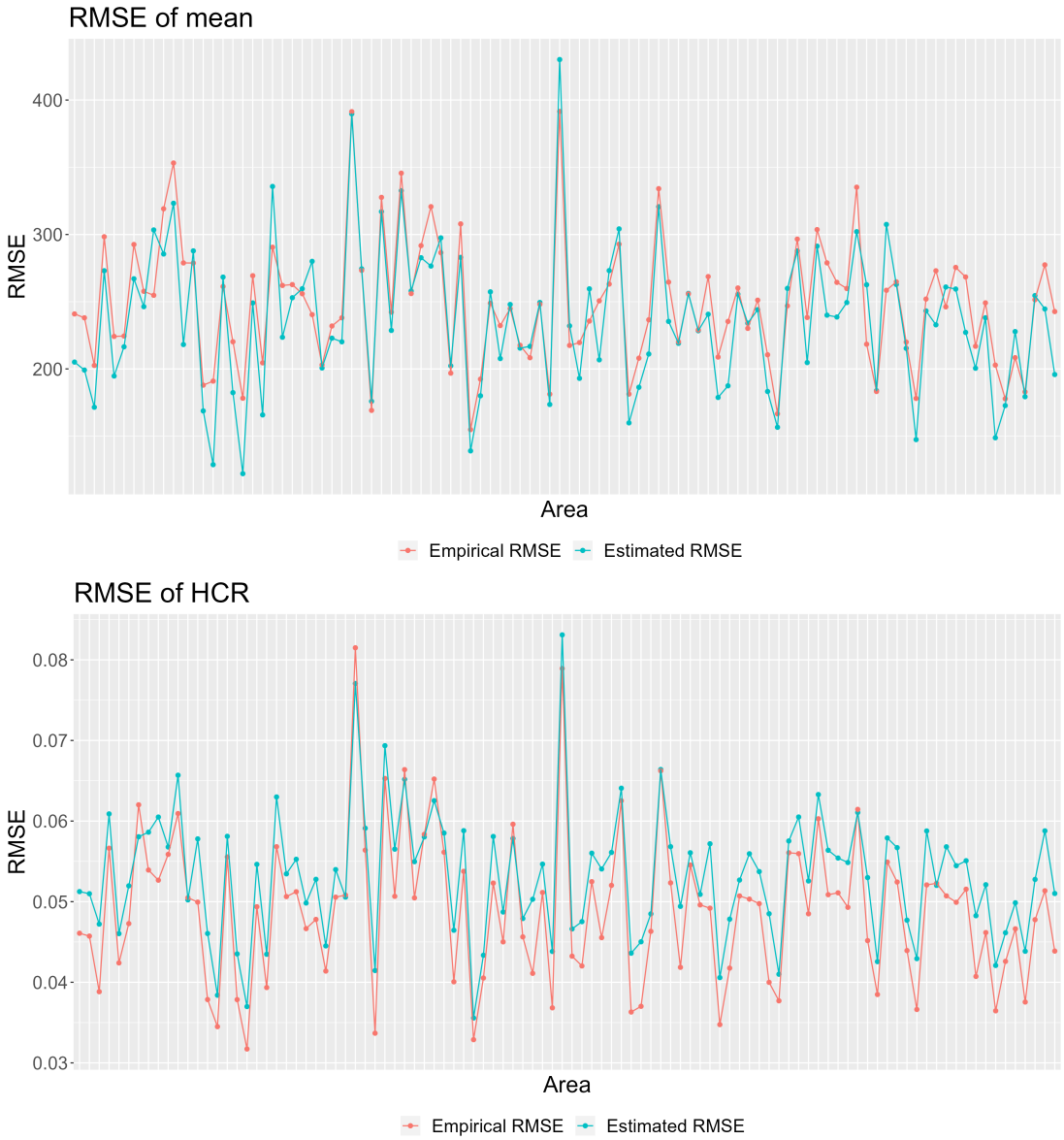


Figure 3.4: Comparison of empirical and estimated RMSE in each area

given datasets.

3.6.1 Direct estimation

For our estimation, we use *IOF 2019/20* as sample data and geospatial data from 2019. Our welfare variable average household consumption per capita per day is taken from the sample data. The covariates for the model-based estimation are obtained from the geospatial auxiliary data including multiple geospatial information. Due to unreliable observations in some grids, we exclude 436 grids from the geospatial auxiliary data. As a result, we have 66803 grids in the geospatial data.

Mozambique is divided into 11 provinces, including 161 districts. In the districts, there are two additional administrative subdivisions. Districts are divided into posts and each post is again divided into localities. Grids are not an administrative division, still, they are geographically smaller than localities. Table 3.4 shows the number of regions at each level in both datasets and the proportions of out-of-sample regions. At district level, we have 8 out-of-sample areas. The more disaggregated a region is, the higher the proportions of out-of-sample regions are.

Table 3.4: Number of regions at each level in *IOF 2019/20* and geospatial auxiliary data and proportion of out-of-sample regions at each level

	District	**Post**	**Locality**	**Grid**
IOF 2019/20	153	366	655	1255
geospatial data	161	456	1274	66803
Proportion of out-of-sample areas	4.97 %	19.74%	48.59%	98.12%

For the initial analysis, we calculate the mean household consumption and the HCR at district level using the HT estimator. The total number of households per district is estimated by the sum of the sampled household weights in each district. The weights of the households are given in the *IOF 2019/20* dataset. The variances of direct estimators are approximately estimated under the assumption that the survey follows a Poisson sampling design. Figure 3.5 shows a map of Mozambique with estimated mean household consumption and HCR. The poverty threshold required to calculate HCR is set at 40.03 MZN, which corresponds to World Bank (2023a). Since the gray districts on the map are not observed in *IOF 2019/20*, it is not possible to get the indicators in these districts. Thus, as shown in Figure 3.6, direct estimates are not reliable in about 50% of districts with coefficient of variation (CV) values higher than 0.2 for both indicators. Due to the out-of-sample districts and unreliable results of direct estimates, an accurate model-based estimation method is essential to obtain precise estimation results at district level.

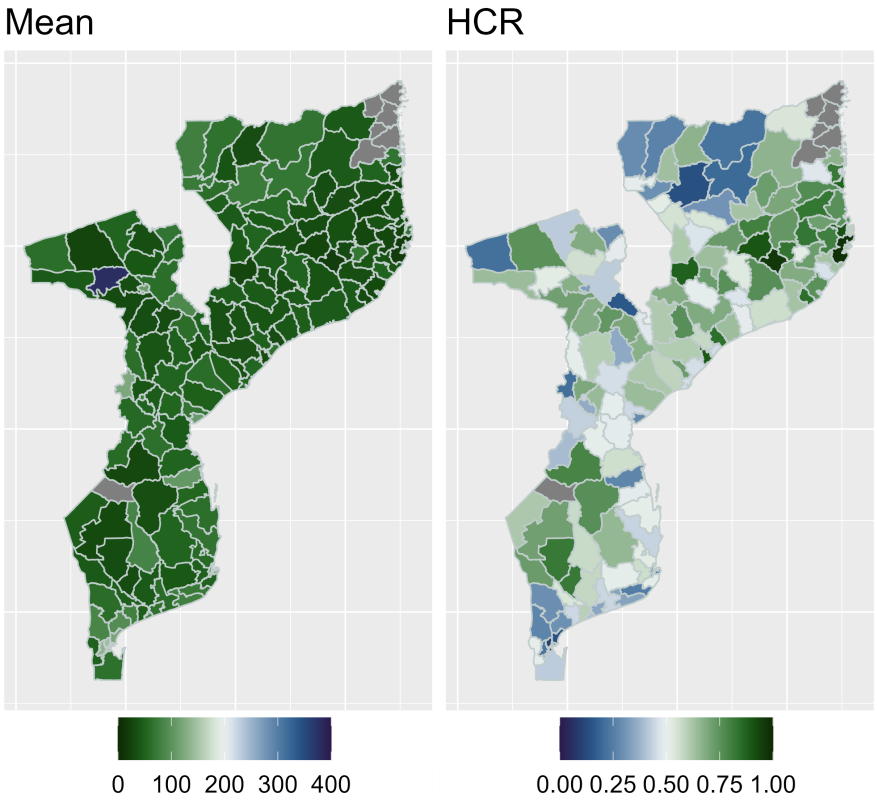


Figure 3.5: Map of direct estimates of mean household consumption and HCR at district level

3.6.2 Model-based estimation

Our available survey data was collected at household level and our geospatial auxiliary data was collected at grid level. We aim to estimate indicators at district level by combining these two datasets. Since

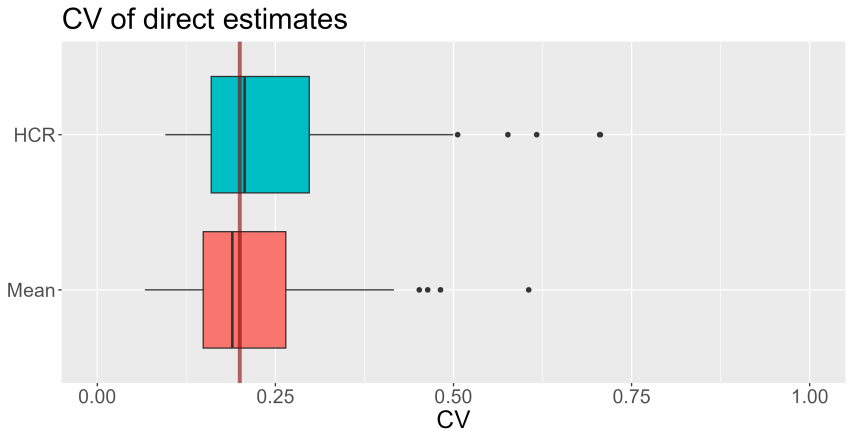


Figure 3.6: Boxplots of CV for the direct estimates

the geospatial covariates are at grid level, the model cannot be defined at a more disaggregated level than the grid. Therefore, we consider the EBLUP of indicators with 4 different models at district level: (a) EBLUP with Fay-Herriot model at district level (*FH*), (b) EBLUP with two-fold model with subarea at post level (*TF_post*), (c) EBLUP with two-fold model with subarea at locality level (*TF_local*), and (d) EBLUP with two-fold model with subarea at grid level (*TF_grid*). The Fay-Herriot model includes only one random effect at district level while the other two-fold models have one random effect at district level and an additional random effect at subarea level. We expect that the two-fold models capture the multilevel data structure better than the Fay-Herriot model by assuming two random effects.

To fit the models, we need direct estimates of the mean and the HCR and explanatory geospatial variables at each level. Geospatial covariates are obtained by taking the average of the geospatial auxiliary variables at each level. We calculate direct estimates by HT estimates based on the *IOF 2019/20*. The corresponding variances are estimated by the Poisson approximation implemented in R package **sae** (Molina and Marhuenda, 2015). The estimated mean household consumption per capita per day at each level is visualized in Figure 3.7. Figure 3.7 shows a strongly skewed distribution of the direct estimates at all levels. This hints at the normality assumptions of Fay-Herriot and two-fold models being violated. Therefore, we apply the logarithmic transformation to the direct estimates of the mean household consumption. Additionally, we apply the arcsine transformation to the square root of the HCR direct estimates so that the predicted values are between 0 and 1. To fit the Fay-Herriot and two-fold model with this transformation, we need the design effect for defining the effective sample size. We calculate the design effect at province level so that the variance estimates of direct estimators are stable enough. The calculated province level design effects are consistently used for all models and effective sample sizes are obtained based on the sample size at each level with the design effect from the province level. To obtain the EBLUP in the original scale, we use the bias corrected back-transformation explained in Section 3.3 for the two-fold models and we use the bias corrected back transformation based on Slud and Maiti (2006) and Sugasawa and Kubokawa (2017) for the Fay-Herriot model implemented in R package **emdi** (Kreutzmann et al., 2019).

With the prepared direct estimates and geospatial covariates data, we perform a variable selection to find an optimal set of explanatory variables with the transformed dependent variable for each model. A fitting way to perform the variable selection for transformed two-fold models is not yet well investigated. Therefore, we use the stepwise selection based on the Akaike information criterion (AIC) (Akaike, 1973) for the Fay-Herriot model in **emdi** for all models (Kreutzmann et al., 2019; Marhuenda et al., 2014). Since the Fay-Herriot model allows only one random effect, each two-fold model is defined at its subarea level and a random effect at the subarea level is included. The chosen optimal set of variables are optimal for the Fay-Herriot model, but suboptimal for the two-fold models because of the missing

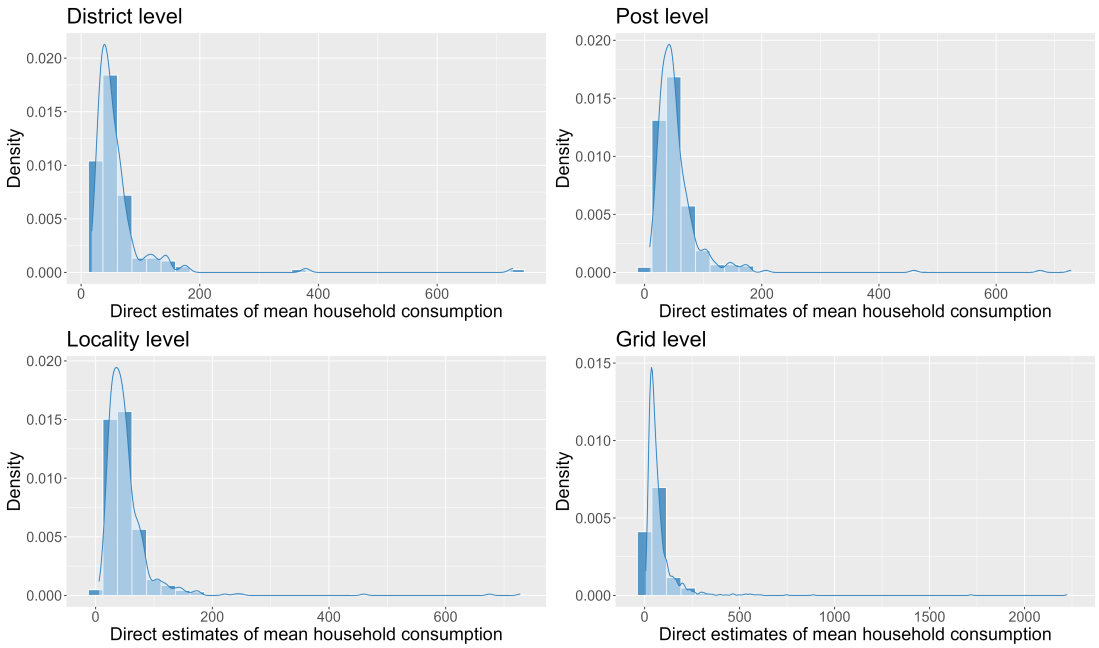


Figure 3.7: Distribution of the HT estimator for the mean household consumption at different regional levels in Mozambique

random effects at area level. We calculate the marginal and conditional R^2 values based on Nakagawa and Schielzeth (2013) to check the predictive power of the chosen optimal models. Table 3.5 shows the R^2 values of the fitted models. The R^2 values of the Fay-Herriot model are generally higher for both indicators compared to the two-fold models. Two-fold models have higher R^2 values for the mean than for the HCR. One interesting point is that the difference between marginal and conditional R^2 for the *TF_grid* is remarkably high for both indicators. This means that the random effects of *TF_grid* explain a high proportion of variation in the data.

In the next step, we examine the normality assumptions of the Fay-Herriot and two-fold models which estimate the mean indicator. Table 3.6 summarizes the skewness, kurtosis, and the p-values of the Shapiro-Wilk normality test (Shapiro and Wilk, 1965) for random effects and residuals. For comparison, we also calculate the measures without logarithmic transformation and these values are shown in brackets below the values with the transformation in the table. Even with the logarithmic transformation, the normality assumptions of residuals are still not fulfilled according to the test results for all models. Furthermore, the normality of subarea level random effects are also rejected for all two-fold models. Still, the normality assumption of district level random effects cannot be rejected for all models and the skewness and kurtosis values of the residuals are much closer to a normal distribution compared to the values without the transformation. Therefore, we conclude that applying logarithmic transformation to the direct estimates of the mean is necessary, since it partially corrects the violation of the normality assumptions.

Based on the fitted models, we calculate the two-fold EBLUP and Fay-Herriot EBLUP. To compare the quality of EBLUP, we estimate the following: RMSE using the proposed method in Section 3.4 for the two-fold EBLUP; analytical RMSE for the mean Fay-Herriot EBLUP based on Slud and Maiti (2006); and bootstrap RMSE for the HCR Fay-Herriot EBLUP by Hadam et al. (2023). Table 3.7 shows the estimated RMSE values of the fitted models. *TF_grid* has the highest RMSE values for both indicators due to the high estimated variance of direct estimators. Especially for the mean, *TF_grid* has remarkably high RMSE values compared to the other models. We suppose that these high RMSE values come from the high and unstable estimated variances of direct estimators. Because of the extreme

Table 3.5: R^2 values of different models used for the estimation of the EBLUP with chosen optimal sets of geospatial covariates

Indicator:		Mean		HCR	
EBLUP	Level of model	Marginal R^2	Conditional R^2	Marginal R^2	Conditional R^2
FH	District	0.5198	0.6779	0.3407	0.5601
TF_post	Post	0.3100	0.4148	0.1591	0.2639
TF_local	Locality	0.2063	0.3105	0.0801	0.2177
TF_grid	Grid	0.3613	0.5801	0.2133	0.5048

Table 3.6: Skewness (S), kurtosis (K), and p-value of Shapiro-Wilk test of random effects and residuals from the fitted models estimating the EBLUP of mean household consumption

EBLUP	Random effects (Area)			Random effects (Subarea)			Residuals		
	S	K	p-value	S	K	p-value	S	K	p-value
FH	0.309	3.365	0.304	-	-	-	1.574	14.863	0.000
	(-0.210)	(3.219)	(0.637)				(7.609)	(65.540)	(0.000)
TF_post	0.275	3.727	0.202	0.507	4.931	0.000	-0.024	5.442	0.000
	(-0.181)	(3.174)	(0.498)	(-0.899)	(4.805)	(0.000)	(8.953)	(101.802)	(0.000)
TF_local	0.226	3.893	0.173	0.339	5.744	0.000	0.018	4.650	0.000
	(-0.162)	(3.624)	(0.344)	(-1.544)	(6.466)	(0.000)	(8.810)	(113.664)	(0.000)
TF_grid	0.047	2.943	0.921	0.182	5.496	0.000	0.241	4.508	0.000
	(0.074)	(3.269)	(0.235)	(-2.807)	(15.634)	(0.000)	(12.364)	(224.019)	(0.000)

values and small sample sizes at grid level, the variance estimates of the direct estimators are very high compared to the other models as shown in Table C2 in the Appendix. *TF_post* and *TF_grid* have smaller RMSE values than *FH* for the mean. *TF_post* has the smallest RMSE values for both indicators, which means that *TF_post* produces the most efficient estimates.

Table 3.7: Summaries of estimated RMSE values

Indicator	EBLUP	Min	1.Q	Median	Mean	3.Q	Max
Mean	*FH*	4.264	6.836	8.565	10.115	10.745	142.566
	TF_post	**4.206**	**6.409**	**7.662**	**9.395**	**9.285**	183.962
	TF_local	4.421	**6.605**	**7.933**	**9.354**	**9.145**	151.634
	TF_grid	4.430	9.702	10.407	$6.50{\cdot}10^{30}$	12.494	$1.05{\cdot}10^{33}$
HCR	FH	0.003	0.065	0.086	0.084	0.103	0.165
	TF_post	**0.001**	**0.065**	**0.085**	**0.083**	**0.100**	**0.158**
	TF_local	0.004	0.075	0.092	0.091	0.110	0.206
	TF_grid	0.021	0.132	0.143	0.137	0.150	0.181

Additionally, we calculated the CV values defined as $CV_i = \sqrt{MSE_i}/\widehat{\theta}_i$ for the estimated indicators where $\widehat{\theta}_i$ denotes the EBLUP. Figure 3.8 shows boxplots of the CV values. For a better illustration, we excluded some outliers higher than 1 from the plots. We can clearly see that *TF_grid* has the highest CV values for both indicators that are affected by the high RMSE values. *TF_local* has smaller CV values for the mean but not for the HCR. Still, the values of CV from *TF_local* are very close to the *FH* for the mean. *TF_post* has the smallest CV for both indicators. Based on the small values of RMSE and CV of *TF_post*, we conclude that *TF_post* is the most appropriate estimation approach of the poverty indicators in Mozambique with the given datasets.

Figure 3.9 shows the estimated mean household consumption per capita per day and HCR values on the map of Mozambique. The estimation results are from the *TF_posto* approach. There is no gray colored district on the map since we can estimate indicators for the out-of-sample districts by using our

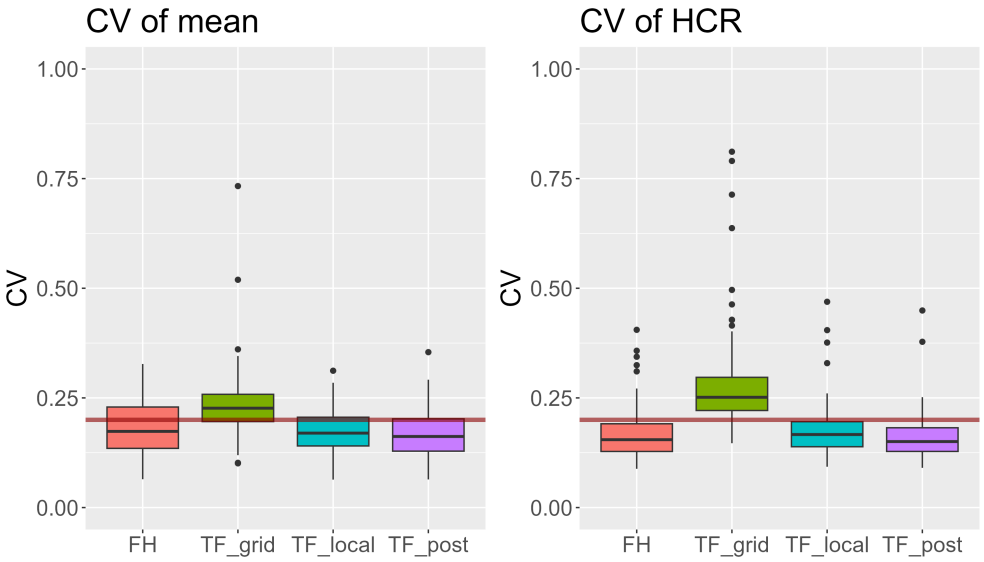


Figure 3.8: Boxplots of CV from different models

model-based estimation method. The northeastern part of Mozambique has lower mean consumption levels and higher HCR values despite the huge economic potential of the region's natural resources. This is most likely a consequence of the violence and instability caused by several insurgencies of political and terrorist groups in the region. Meanwhile, the districts with higher mean consumption are located in the red box on the left side of Figure 3.9. These districts belong to the capital city of Mozambique, Maputo City. Evidently, the concentration of wealth and power in Maputo city contributes to the comparatively high mean consumption and low HCR.

3.7 Conclusion

In the present study, we estimate poverty indicators in Mozambique at district level by using geospatial data as an alternative auxiliary data source in transformed two-fold models. Usually, researchers use household survey data and auxiliary census data in SAE models. However, it is difficult to get access to up to date census data. Therefore, we recommend the use of geospatial data as alternative census data. Geospatial data is up to date, easy to acquire, and covers all regions of interest. We have a multilevel data structure in our household survey and geospatial data, while standard SAE models only have two levels: areas and households. We suggest the use of a two-fold model with geospatial data because this model can capture the hierarchical regional structure more precisely by including two random effects defined at two regional levels. We extend the two-fold model with transformations and introduce a bias corrected back-transformation. Moreover, we propose a MSE estimation method for the two-fold EBLUP. Our model-based simulation study shows that the proposed transformed two-fold EBLUP with bias corrected transformation has clear advantages over the Fay-Herriot EBLUP. Furthermore, the MSE estimation method for the two-fold EBLUP also performs well. In the application, we compare the estimates of different two-fold models at different subarea levels and the Fay-Herriot model. We see that there are differences between two-fold models depending on the subarea level, and that a two-fold model at a stable subarea level (e.g., *TF_posto* in our application) produces more efficient estimates.

There are some discussion points for further research. First, it is not yet clear how exactly the subarea should be defined. We suppose that it depends on the variance estimates of the direct estimators. We observe in the application that the *TF_grid* produces much worse estimators compared to the *FH* due

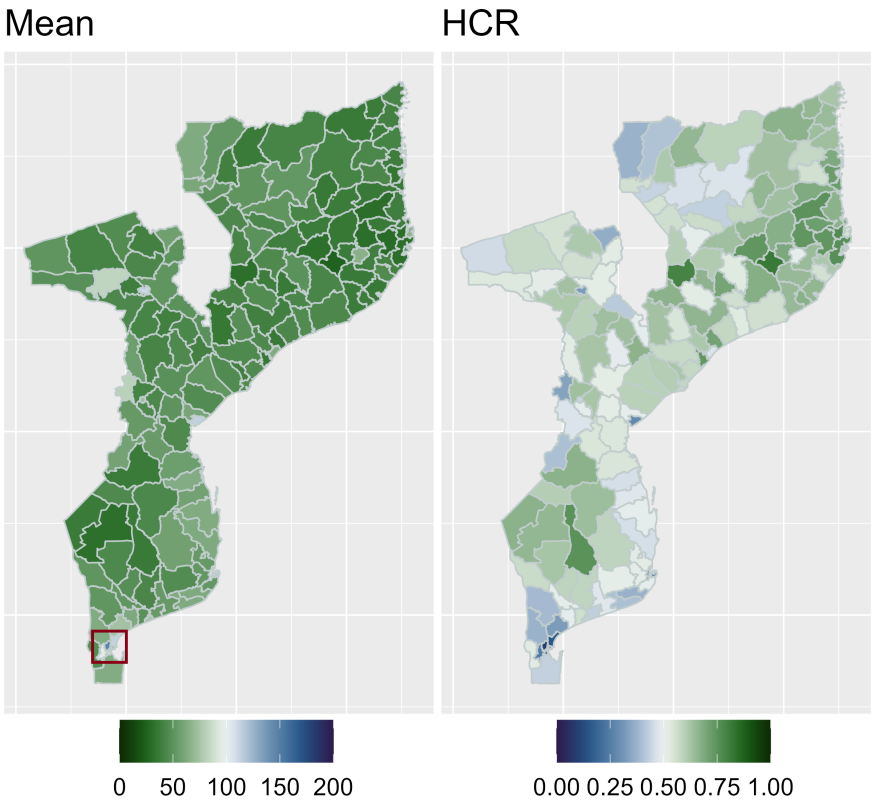


Figure 3.9: Map of the estimated mean household consumption and HCR in Mozambique

to the highly unstable variances of the direct estimates. Additionally, we assume the advantage of our two-fold model over the Fay-Herriot model is related to the variances of the random effects. When the variance of the subarea level random effect is not large enough compared to the variance of the area level random effect, the advantage may be smaller. Consequently, we suppose that the subarea level must be defined at a level where we can have a stable variance estimation of direct estimates and a high variance of the subarea level random effect. Second, data driven transformations may also be possible for the two-fold model. The logarithmic transformation frequently cannot correct the violation in normality assumptions completely. An appropriate data driven transformation such as Box-Cox transformation (Box and Cox, 1964) or dual power transformation (Yang, 2006) can improve the performance of transformations correcting the violated distributional assumptions (Rojas-Perilla et al., 2020; Sugasawa and Kubokawa, 2015). Third, a valid variable selection method for the two-fold model is necessary. The selection criteria for the Fay-Herriot model by Marhuenda et al. (2014) could be a good starting point. When the data driven transformation is expected, an extension of the Jacobian adjusted conditional AIC by Lee et al. (2023) for the two-fold model could also be possible.

Acknowledgments

Lee gratefully acknowledges support by a scholarship of Studienstiftung des deutschen Volkes.

Appendix C

C.1 Graphics and tables

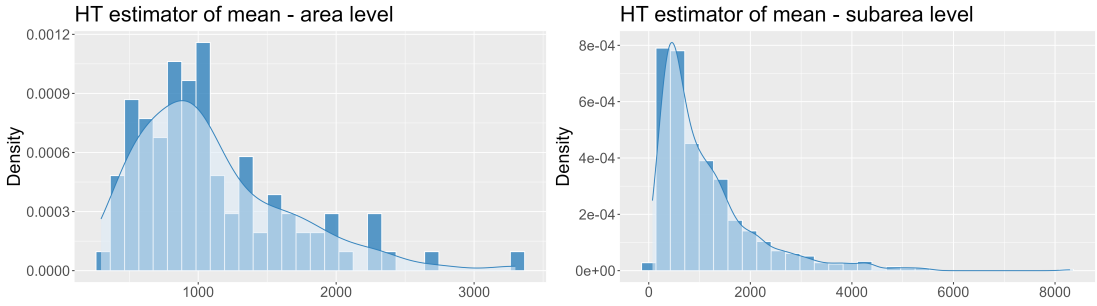


Figure C1: Distribution of the HT estimator for the mean at area and grid level from 200th simulation sample dataset

Table C1: Skewness and kurtosis of the HT estimator for the mean at area and subarea level of 500 simulation datasets

Level	Moment	Min	1.Q	Median	Mean	3.Q	Max
Area	Skewness	0.212	0.987	1.230	1.342	1.593	3.200
	Kurtosis	2.172	3.850	4.830	5.800	6.689	19.930
Subarea	Skewness	1.315	2.209	2.561	2.840	3.157	10.709
	Kurtosis	5.153	10.193	12.970	17.965	18.392	197.890

Table C2: Summaries of estimated variances for the direct estimates of the mean

Indicator	Level of direct estimates	Min	1.Q	Median	Mean	3.Q	Max
Mean	District	12.17	46.12	88.29	399.77	178.32	18211.44
	Post	11.97	84.22	168.99	644.49	336.73	61101.05
	Locality	9.94	118.09	236.28	768.69	466.77	61101.05
	Grid	9.9	186.0	404.2	3498.1	1055.0	843992.8

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), Proceedings of the 2nd International Symposium on Information Theory, pp. 267–281. Akademiai Kiado.

Akaike, H. (1978). On the likelihood of a time series model. Journal of the Royal Statistical Society: Series D (The Statistician) 27(3/4), 217–235.

Alfons, A. and M. Templ (2013). Estimation of social exclusion indicators from complex surveys: The R package **laeken**. Journal of Statistical Software 54(15), 1–25.

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An error-component model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association 83 (401), 28–36.

Benavent, R. and D. Morales (2016). Multivariate Fay-Herriot models for small area estimation. Computational Statistics & Data Analysis 94, 372–390.

BLS (2023a). Handbook of methods. Consumer price index. Calculation. https://www.bls.gov/opub/hom/cpi/calculation.htm. Accessed: 30.04.2023.

BLS (2023b). Handbook of methods. Consumer price index. Presentation. https://www.bls.gov/opub/hom/cpi/presentation.htm. Accessed: 28.04.2023.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological) 26(2), 211–252.

Bunke, O., B. Droge, and J. Polzehl (1999). Model selection, transformations and variance estimation in nonlinear regression. Statistics: A Journal of Theoretical and Applied Statistics 33(3), 197–240.

Burgard, J. P., M. D. Esteban, D. Morales, and A. Pérez (2020). A Fay-Herriot model when auxiliary variables are measured with error. TEST 29(1), 166–195.

Burnham, K. P. and D. R. Anderson (2002). Model Selection and Multimodel Inference (2nd ed.). New York: Springer.

Carter, G. and J. Rolph (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. Journal of the American Statistical Association 69(348), 880–885.

Corral, P., K. Himelein, K. McGee, and I. Molina (2021). A map of the poor or a poor map? Mathematics 9(21), 2780.

Dawber, J., N. Würz, P. A. Smith, T. Flower, H. Thomas, T. Schmid, and N. Tzavidis (2022). Experimental UK regional consumer price inflation with model-based expenditure weights. Journal of Official Statistics 38(1), 213–237.

Donohue, M. C., R. Overholser, R. Xu, and F. Vaida (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. Biometrika 98(3), 685–700.

Duarte, M. and A. L. Wolman (2008). Fiscal policy and regional inflation in a currency union. Journal of International Economics 74(2), 384–401.

Dutot, N. (1738). Réflexions politiques sur les finances et le commerce. The Hague: Chez les Freres Vaillant & Nicolas Prevost.

Erciulescu, A. L., N. B. Cruze, and B. Nandram (2019). Model-based county level crop estimates incorporating auxiliary sources of information. Journal of the Royal Statistical Society: Series A (Statistics in Society) 182(1), 283–303.

Esteban, M. D., M. J. Lombardía, E. López-Vizcaíno, D. Morales, and A. Pérez (2020). Small area estimation of proportions under area-level compositional mixed models. TEST 29(3), 793–818.

Fang, Y. (2011). Asymptotic equivalence between cross-validations and Akaike information criteria in mixed-effects models. Journal of Data Science 9(1), 15–21.

Fay, R. E. and R. A. Herriot (1979). Estimation of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association 74 (366), 269–277.

Foster, J., J. Greer, and E. Thorbecke (1984). A class of decomposable poverty measures. Econometrica 52(3), 761–766.

Friedl, M. and D. Sulla-Menashe (2022). MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MCD12Q1.061.

Gini, C. (1912). Variabilitá e mutuabilitá: contributo allo studio delle distribuzioni e delle relazioni statistiche. Bologna: Tipografia di Paolo Cuppini.

González-Manteiga, W., M. J. Lombardía, I. Molina, D. Morales, and L. Santamaría (2008). Bootstrap mean squared error of a small-area EBLUP. Journal of Statistical Computation and Simulation 78(5), 443–462.

Greven, S. and T. Kneib (2010). On the behaviour of marginal and conditional AIC in linear mixed models. Biometrika 97(4), 773–789.

Gurka, M. J., L. J. Edwards, K. E. Muller, and L. L. Kupper (2006). Extending the Box-Cox transformation to the linear mixed model. Journal of the Royal Statistical Society: Series A (Statistics in Society) 169(2), 273–288.

Hadam, S., N. Würz, A.-K. Kreutzmann, and T. Schmid (2023). Estimating regional unemployment with mobile network data for functional urban areas in Germany. Statistical Methods & Applications. forthcoming.

Han, B. (2013). Conditional Akaike information criterion in the Fay-Herriot model. Statistical Methodology 11, 53–67.

Harmening, S., A.-K. Kreutzmann, S. Schmidt, N. Salvati, and T. Schmid (2023). A framework for producing small area estimates based on area-level models in R. The R Journal 15(1), 316–341.

Hodges, J. S. and D. J. Sargent (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. Biometrika 88(2), 367–379.

Hoeting, J. A. and J. G. Ibrahim (1998). Bayesian predictive simultaneous variable and transformation selection in the linear model. Computational Statistics & Data Analysis 28(1), 87–103.

Hoeting, J. A., A. E. Raftery, and D. Madigan (2002). Bayesian variable and transformation selection in linear regression. Journal of Computational and Graphical Statistics 11(3), 485–507.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47(260), 663–685.

Huffman, G. J., E. F. Stocker, D. T. Bolvin, E. J. Nelkin, and J. Tan (2019). GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V06, Greenbelt, MD [Data set]. Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/GPM/IMERG/3B-MONTH/06.

IMF (2023). World economic outlook (October 2023). https://www.imf.org/external/datamapper/datasets/WEO. Accessed: 23.01.2024.

International Labour Organization (2024). Household income and expenditure surveys. https://www.ilo.org/surveyLib/index.php/catalog/HIES/?page=1&ps=15&repo=HIES. Accessed: 22.04.2024.

Jiang, J. (2019). Robust Mixed Model Analysis. Singapore: World Scientific.

Jiang, J. and P. Lahiri (2006). Mixed model prediction and small area estimation. TEST 15(1), 1–96.

Jiang, J., P. Lahiri, S.-M. Wan, and C.-H. Wu (2001). Jackknifing in the Fay-Herriot model with an example. In Proceedings of the Seminar on Funding Opportunity in Survey Research, pp. 75–97. Federal Committee on Statistical Methodology.

Kosfeld, R., H.-F. Eckey, and M. Schüßler (2009). Ökonometrische Messung regionaler Preisniveaus auf der Basis örtlich beschränkter Erhebungen. (RatSWD research notes, no. 33). Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://www.konsortswd.de/publikation/rn33-2009/. Accessed: 14.05.2024.

Krause, J., J. P. Burgard, and D. Morales (2022). Robust prediction of domain compositions from uncertain data using isometric logratio transformations in a penalized multivariate Fay-Herriot model. Statistica Neerlandica 76(1), 65–96.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2019). The R package **emdi** for estimating and mapping regionally disaggregated indicators. Journal of Statistical Software 91(7), 1–33.

Lachos, V. H., D. K. Dey, and V. G. Cancho (2009). Robust linear mixed models with skew-normal independent distributions from a Bayesian perspective. Journal of Statistical Planning and Inference 139(12), 4098–4110.

Lee, Y., N. Rojas-Perilla, M. Runge, and T. Schmid (2023). Variable selection using conditional AIC for linear mixed models with data-driven transformations. Statistics and Computing 33(27), 1–17.

Liang, H., H. Wu, and G. Zou (2008). A note on conditional AIC for linear mixed-effects models. Biometrika 95(3), 773–778.

Marchetti, S. and L. Secondi (2017). Estimates of household consumption expenditure at provincial level in Italy by using small area estimation methods: "Real" comparisons using purchasing power parities. Social Indicators Research 131(1), 215–234.

Marhuenda, Y., D. Morales, and M. del Carmen Pardo (2014). Information criteria for Fay-Herriot model selection. Computational Statistics & Data Analysis 70, 268–280.

Masaki, T., D. Newhouse, A. R. Silwal, A. Bedada, , and R. Engstrom (2022). Small area estimation of non-monetary poverty with geospatial data. Statistical Journal of the IAOS 38(3), 1035–1051.

Molina, I. and Y. Marhuenda (2015). **sae**: An R package for small area estimation. The R Journal 7(1), 81–98.

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

Morales, D., M. D. Esteban, A. Pérez, and T. Hobza (2021). A Course on Small Area Estimation and Mixed Models (1st ed.). Cham: Springer.

Müller, S., J. L. Scealy, and A. H. Welsh (2013). Model selection in linear mixed models. Statistical Science 28(2), 135–167.

Nagayasu, J. (2011). Heterogeneity and convergence of regional inflation (prices). Journal of Macroeconomics 33(4), 711–723.

Nakagawa, S. and H. Schielzeth (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4(2), 133–142.

Neves, A., D. Silva, and S. Correa (2013). Small domain estimation for the Brazilian service sector survey. Estadística 65(185), 13–37.

Newhouse, D. (2023). Small area estimation of poverty and wealth using geospatial data. What have we learned so far? (Policy research working paper 10512). World Bank Group. http://hdl.handle.net/10986/40028. Accessed: 08.03.2024.

Newhouse, D., J. D. Merfeld, A. P. Ramakrishnan, T. Swartz, and P. Lahiri (2022). Small area estimation of monetary poverty in Mexico using satellite imagery and machine learning. (Policy research working paper 10175). World Bank Group. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099430309142231728/idu0660868530404c0414e0bf180797b525682a5. Accessed: 08.03.2024.

Office of International Religious Freedom (2022). 2022 Report on international religious freedom: Lebanon. U.S. department of state. https://www.state.gov/reports/2022-report-on-international-religious-freedom/lebanon/. Accessed: 19.03.2024.

Permatasari, N. and A. Ubaidillah (2021). **msae**: An R package of multivariate Fay-Herriot models for small area estimation. The R Journal 13(2), 111–122.

Pfeffermann, D. (2013). New important developments in small area estimation. Statistical Science 28 (1), 40–68.

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

Pusponegoro, N. H., A. Kurnia, K. A. Notodiputro, A. M. Soleh, and E. T. Astuti (2021). Area specific effects selection of small area estimation for construction of regional consumer price indices in Indonesia. Communications in Mathematical Biology and Neuroscience 2021, 90.

Raghunathan, T. E., D. Xie, N. Schenker, V. L. Parsons, W. W. Davis, K. W. Dodd, and E. J. Feuer (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. Journal of the American Statistical Association 102(478), 474–486.

Rao, J. N. K. and I. Molina (2015). Small Area Estimation (2nd ed.). Hoboken: John Wiley & Sons.

Rao, J. N. K. and M. Yu (1994). Small-area estimation by combining time-series and cross-sectional data. The Canadian Journal of Statistics 22(4), 511–528.

R Core Team (2021). R: A language and environment for statistical computing. Vienna: R foundation for statistical computing.

R Core Team (2022). R: A language and environment for statistical computing. Vienna: R foundation for statistical computing.

RDC (2018a). Einkommens- und Verbrauchsstichprobe (NGT) 2018, Scientific-Use-File (SUF) [Data set]. https://doi.org/10.21242/63231.2018.00.00.3.1.0.

RDC (2018b). Verbraucherpreisindex für Deutschland 2018, Scientific Use File (SUF) [Data set]. https://doi.org/10.21242/61111.2018.00.00.3.1.0.

RDC (2021). Metadatenreport. Teil I: Allgemeine und methodische Informationen zu den Einzeldaten des Verbraucherpreisindex (EVAS-Nummer: 61111). https://www.forschungsdatenzentrum.de/sites/default/files/vpi_2005-2020_mdr_teil1_statistik_v2.pdf. Accessed: 18.04.2023.

Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis (2020). Data-driven transformations in small area estimation. Journal of the Royal Statistical Society: Series A (Statistics in Society) 183(1), 121–148.

Román, M. O., Z. Wang, Q. Sun, V. Kalb, S. D. Miller, A. Molthan, L. Schultz, J. Bell, E. C. Stokes, B. Pandey, K. C. Seto, D. Hall, T. Oda, R. E. Wolfe, G. Lin, N. Golpayegani, S. Devadiga, C. Davidson, S. Sarkar, C. Praderas, J. Schmaltz, R. Boller, J. Stevens, O. M. R. González, E. Padilla, J. Alonso, Y. Detrés, R. Armstrong, I. Miranda, Y. Conte, N. Marrero, K. MacManus, T. Esch, and E. J. Masuoka (2018). NASA's black marble nighttime lights product suite. Remote Sensing of Environment 210, 113–143.

Rosa, G. J. M., C. Padovani, and D. Gianola (2003). Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. Biometrical Journal 45(5), 573–590.

SBJ (2022). 2020-Base explanation of the consumer price index. https://www.stat.go.jp/english/data/cpi/1590.html. Accessed: 30.04.2023.

Scealy, J. L. and A. H. Welsh (2017). A directional mixed effects model for compositional expenditure data. Journal of the American Statistical Association 112(517), 24–36.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: Estimating literacy rates in Senegal. Journal of the Royal Statistical Society: Series A (Statistics in Society) 180(4), 1163–1190.

Shang, J. and J. E. Cavanaugh (2008). Bootstrap variants of the Akaike information criterion for mixed model selection. Computational Statistics & Data Analysis 52(4), 2004–2021.

Shapiro, S. S. and M. Wilk (1965). An analysis of variance test for normality (complete samp1es). Biometrika 52(3/4), 591–611.

Sinha, S. K. and J. N. K. Rao (2009). Robust small area estimation. The Canadian Journal of Statistics 37(3), 381–399.

Slud, E. V. and T. Maiti (2006). Mean-squared error estimation in transformed Fay-Herriot models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(2), 239–257.

Smith, P. A. (2021). Estimating sampling errors in consumer price indices. International Statistical Review 89(3), 481–504.

Statistics Canada (2023). The Canadian consumer price index reference paper. https://www150.statcan.gc.ca/n1/pub/62-553-x/62-553-x2015001-eng.pdf. Accessed: 30.04.2023.

Statistisches Bundesamt (2013). Das Systematische Verzeichnis der Einnahmen und Ausgaben der privaten Haushalte, Ausgabe 2013. https://www.destatis.de/DE/Methoden/Klassifikationen/Private-Haushalte/sea-2013.pdf. Accessed: 18.04.2023.

Statistisches Bundesamt (2018). Privathaushalte: Bundesländer, Jahre (bis 2019). https://www-genesis.destatis.de/genesis/online?operation=previous&levelindex=2&step=2&titel=Ergebnis&levelid=1691420348856&acceptscookies=false#abreadcrumb. Accessed: 24.03.2023.

Statistisches Bundesamt (2021a). Qualitätsbericht - Einkommens- und Verbrauchsstichprobe. https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Einkommen-Konsum-Lebensbedingungen/einkommens-verbrauchsstichprobe-2018.pdf. Accessed: 18.04.2023.

Statistisches Bundesamt (2021b). Qualitätsbericht - Preise. Verbraucherpreisindex für Duetschland. https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Preise/verbraucherpreis.pdf. Accessed: 18.04.2023.

Statistisches Bundesamt (2022). Wirtschaftsrechnungen - Einkommens- und Verbrauchsstichprobe. Aufgabe, Methode und Durchführung. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Einkommen-Einnahmen-Ausgaben/Publikationen/Downloads-Einkommen/evs-aufgabe-methode-durchfuehrung-2152607189004.pdf. Accessed: 18.04.2023.

Statistisches Bundesamt (2023). Hintergrundpapier zur Revision des Verbraucherpreisindex für Deutschland 2023. https://www.destatis.de/DE/Themen/Wirtschaft/Preise/Verbraucherpreisindex/Methoden/Downloads/Hintergrundpapier-VPI-Revision_2020.pdf. Accessed: 31.04.2023.

Steele, J. E., P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y.-A. de Montjoye, A. M. Iqbal, K. N. Hadiuzzaman, X. Lu, E. Wetter, A. J. Tatem, and L. Bengtsson (2017). Mapping poverty using mobile phone and satellite data. Journal of the Royal Society: Interface 14(127), 20160690.

Sugasawa, S. and T. Kubokawa (2015). Parametric transformed Fay-Herriot model for small area estimation. Journal of Multivariate Analysis 139, 295–311.

Sugasawa, S. and T. Kubokawa (2017). Transforming response values in small area prediction. Computational Statistics & Data Analysis 114, 47–60.

The DHS Program (2023). Demographic and health survey (DHS). https://dhsprogram.com/Methodology/Survey-Types/DHS.cfm. Accessed: 22.04.2024.

Torabi, M. and J. N. K. Rao (2014). On small area estimation under a sub-area level model. Journal of Multivariate Analysis 127, 36–55.

Tzavidis, N., L.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla (2018). From start to finish: a framework for the production of small area official statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society) 181(4), 927–979.

United Nations (2021). Guide on producing CPI under lockdown. https://unece.org/sites/default/files/2021-12/Guide%20on%20producing%20CPI%20under%20lockdown%20WEB%20version.pdf. Accessed: 31.04.2023.

United Nations (2023). The sustainable development goals report 2023: Special edition. https://unstats.un.org/sdgs/report/2023/. Accessed: 26.03.2024.

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

Verbeke, G. and E. Lesaffre (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. Computational Statistics & Data Analysis 23(4), 541–556.

Verbeke, G. and G. Molenberghs (2000). Linear mixed models for longitudinal data (1st ed.). New York: Springer.

Weinand, S. and L. von Auer (2020). Anatomy of regional price differentials: evidence from micro-price data. Spatial Economic Analysis 15(4), 413–440.

World Bank (2016). Facing hard choices: Mozambique economic update. https://www.worldbank.org/en/news/press-release/2016/12/12/facing-hard-choices-mozambique-economic-update. Accessed: 23.01.2024.

World Bank (2023a). Mozambique poverty assessment, June 2023: Poverty reduction setback in times of compounding shocks. https://doi.org/10.1596/40106. Accessed: 04.03.2024.

World Bank (2023b). The World Bank in Mozambique. https://www.worldbank.org/en/country/mozambique/overview. Accessed: 23.01.2024.

Yang, Z. (2006). A modified family of power transformations. Economics Letters 92(1), 14–19.

Ybarra, L. M. R. and S. L. Lohr (2008). Small area estimation when auxiliary information is measured with error. Biometrika 95(4), 919–931.

Yeh, C., A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. Nature Communications 11, 2583.

Zhang, D. and M. Davidian (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. Biometrics 57(3), 795–802.