Strategies for Multiply Imputed Survey Data and Modeling in the Context of Small Area Estimation

Inaugural-Dissertation zur Erlangung des akademischen Grades eines Doktors/einer Doktorin der Wirtschaftswissenschaft des Fachbereichs Wirtschaftswissenschaft der Freien Universität Berlin

vorgelegt von
Marina Runge
Berlin, 2023

Marina Runge, *Strategies for Multiply Imputed Survey Data and Modeling in the Context of Small Area Estimation,*
February 2023

Supervisors:
Prof. Dr. Timo Schmid (Otto-Friedrich-Universität Bamberg)
Prof. Nicola Salvati, Ph.D. (Università di Pisa)
Location:
Berlin

Date of defense:
June 27, 2023

Acknowledgements

I would like to express my deepest gratitude and respect to my supervisor, Prof. Dr. Timo Schmid (Otto-Friedrich-Universität Bamberg, Germany) for his support, guidance and encouragement. His mentoring and fruitful discussions were of great value to the success of this thesis.

For his scientific expertise and profound statistical input, I am grateful to Prof. Nicola Salvati, PhD (Università di Pisa, Italy).

I would like to thank my colleagues at the Chair of Statistics and the Statistical Consulting Unit fu:stat for the pleasant and supportive working environment and friendship.

I appreciate gratefully the support of the German Research Foundation within the TESAP project (281573942).

I am very grateful to all those who have accompanied me on my way to this dissertation, especially my friends and family, who give me joy and strength and whose value to me is inestimable.

Publication List

The publications listed below are the result of the research carried out in this thesis titled, "Strategies for Multiply Imputed Survey Data and Modeling in the Context of Small Area Estimation."

1. Kreutzmann, A.-K., Marek, P., Runge, M., Salvati, N. and Schmid, T. (2022) **The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany**, *Journal of Applied Statistics, 49(13), pp. 3278-3299,* doi: https://doi.org/10.1080/02664763.2021.1941805. Accepted and published.

2. Runge, M. and Schmid, T. (2023) **Small area estimation with multiply imputed survey data,** *Journal of Official Statistics, 39(4), pp. 507-533,* doi: https://doi.org/10.2478/jos-2023-0024. Accepted and published.

3. Lee, Y., Rojas-Perilla, N., Runge, M. and Schmid, T. (2023) **Variable selection using conditional AIC for linear mixed models with data-driven transformations.** *Statistics and Computing 33(27).* doi: https://doi.org/10.1007/s11222-022-10198-9. Accepted and published.

4. Harmening, S., Runge, M. and Schmid, T. (2023) **Area-level small area estimation with random forests**. *Working Paper*.

5. Runge, M. (2023) **Estimating intra-regional inequality with an application to German spatial planning regions.** *Journal of Official Statistics, 39(2), pp.203-228.* doi: https://doi.org/10.2478/jos-2023-0010. Accepted and published.

Contents

Introduction

To target resources and policies where they are most needed, it is essential that policy-makers are provided with reliable socio-demographic indicators on sub-groups. These sub-groups can be defined by regional divisions or by demographic characteristics and are referred to as areas or domains. Information on these domains is usually obtained through surveys, often planned at a higher level, such as the national level. As sample sizes at disaggregated levels may become small or unavailable, estimates based on survey data alone may no longer be considered reliable or may not be available. Increasing the sample size is time consuming and costly. Small area estimation (SAE) methods aim to solve this problem and achieve higher precision. SAE methods enrich information from survey data with data from additional sources and "borrow" strength from other domains (Rao and Molina, 2015; Tzavidis et al., 2018). This is done by modeling and linking the survey data with administrative or register data and by using area-specific structures. Auxiliary data are traditionally population data available at the micro or aggregate level that can be used to estimate unit-level models (Battese et al., 1988; Molina and Rao, 2010) or area-level models (Fay and Herriot, 1979). Due to strict privacy regulations, it is often difficult to obtain these data at the micro level. Therefore, models based on aggregated auxiliary information, such as the Fay-Herriot model and its extensions, are of great interest for obtaining SAE estimators.

Despite the problem of small sample sizes at the disaggregated level, surveys often suffer from high non-response. One possible solution to item non-response is multiple imputation (MI), which replaces missing values with multiple plausible values. The missing values and their replacement introduce additional uncertainty into the estimate. Part I focuses on the Fay-Herriot model, where the resulting estimator is a combination of a design-unbiased estimator based only on the survey data (hereafter called the direct estimator) and a synthetic regression component. Solutions are presented to account for the uncertainty introduced by missing values in the SAE estimator using Rubin's rules (Rubin, 1987). Since financial assets and wealth are sensitive topics, surveys on this type of data suffer particularly from item non-response. Chapter 1 focuses on estimating private wealth at the regionally disaggregated level in Germany. Data from the 2010 Household Finance and Consumption Survey (HFCS) (Household Finance and Consumption Network, 2016b) are used for this application. In addition to the non-response problem, income and wealth data are often right-skewed, requiring a transformation to fully satisfy the normality assumptions of the model. Therefore, Chapter 1 presents a modified Fay-Herriot approach that incorporates the uncertainty of missing values into the log-transformed direct estimator of a mean. Chapter 2 complements Chapter 1 by presenting a framework that extends the general class of transformed Fay-Herriot models to account for the additional un-

certainty due to MI by including it in the direct component and simultaneously in the regression component of the Fay-Herriot estimator. In addition, the uncertainty due to missing values is also included in the mean squared error estimator, which serves as the uncertainty measure. The estimation of a mean, the use of the log transformation for skewed data, and the arcsine transformation for proportions as target indicators are considered. The proposed framework is evaluated for the three cases in a model-based simulation study. To illustrate the methodology, 2017 data from the HFCS (Household Finance and Consumption Network, 2020a) for European Union countries are used to estimate the average value of bonds at the national level. The approaches presented in Chapters 1 and 2 contribute to the literature by providing solutions for estimating SAE models in the presence of multiply imputed survey data. In particular, Chapter 2 presents a general approach that can be extended to other indicators.

To obtain the best possible SAE estimator in terms of accuracy and precision, it is important to find the optimal model for the relationship between the target variable and the auxiliary data. The notion of "optimal" can be multifaceted. One way to look at optimality is to find the best transformation of the target variable to fully satisfy model assumptions or to account for nonlinearity. Another perspective is to identify the most important covariates and their relationship to each other and to the target variable. Part II of this dissertation therefore brings together research on optimal transformations and model selection in the context of SAE. Chapter 3 considers both problems simultaneously for linear mixed models (LMM) and proposes a model selection approach for LMM with data-driven transformations. In particular, the conditional Akaike information criterion (Vaida and Blanchard, 2005) is adapted by introducing the Jacobian into the criterion to allow comparison of models at different scales. The methodology is evaluated in a simulation experiment comparing different transformations with different underlying true models. Since SAE models are LMMs, this methodology is applied to the unit-level small-area method, the empirical best predictor (EBP) (Molina and Rao, 2010), in an application with Mexican survey and census data (ENIGH - National Survey of Household Income and Expenditure) and shows improvements in efficiency when the optimal (linear mixed) model and the transformation parameters are found simultaneously. Chapter 3 bridges the gap between model selection and optimal transformations to satisfy normality assumptions in unit-level SAE models in particular and LMMs in general. Chapter 4 explores the problem of model selection from a different perspective and for area-level data. To model interactions between auxiliary variables and nonlinear relationships between them and the dependent variable, machine learning methods can be a versatile tool. For unit-level SAE models, mixed-effects random forests (MERFs) (Hajjem et al., 2014; Krennmair and Schmid, 2022) provide a flexible solution to account for interactions and nonlinear relationships, ensure robustness to outliers, and perform implicit model selection. In Chapter 4, the idea of MERFs is transferred to area-level models and the linear regression synthetic part of the Fay-Herriot model is replaced by a random forest (Breiman, 2001) to benefit from the above properties and to provide an alternative modeling approach. Chapter 4 therefore contributes to the literature by proposing a first way to combine area-level SAE models with random forests for mean estimation to allow for interactions, nonlinear relationships, and implicit variable selection. Another advantage of random forest is its non-extrapolation property, i.e. the range of predictions is limited by the lowest

and highest observed values. This could help to avoid transformations at the area-level when estimating indicators defined in a fixed range. The standard Fay-Herriot model was originally developed to estimate a mean, and transformations are required when the indicator of interest is, for example, a share or a Gini coefficient. This usually requires the development of appropriate back-transformations and MSE estimators. Chapter 5 presents a Fay-Herriot model for estimating logit-transformed Gini coefficients with a bias-corrected back-transformation and a bootstrap MSE estimator. A model-based simulation is performed to show the validity of the methodology, and regionally disaggregated data from Germany (Socio-Economic Panel, 2019) are used to illustrate the proposed approach. Chapter 5 contributes to the existing literature by providing, from a frequentist perspective, an alternative to the Bayesian area-level model for estimating Gini coefficients using a logit transformation (Fabrizi and Trivisano, 2016).

Part I

Estimation of Area-Level Models with Multiply Imputed Survey Data

Chapter 1

The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany

This is the peer reviewed version of the following article: Kreutzmann, A.-K., Marek, P., Runge, M., Salvati, N. and Schmid, T. (2022) The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany, *Journal of Applied Statistics, 49(13), pp. 3278-3299,* which has been published in final form at https://doi.org/10.1080/02664763.2021.1941805. Due to copyright requirements, this article has been excluded and can be accessed at https://doi.org/10.1080/02664763.2021.1941805.

Chapter 2

Small area estimation with multiply imputed survey data

2.1 Motivation

Financial reports based on asset data can provide insights into a wide range of issues of major importance for political decisions and can help in the precise allocation of funds. In addition, wealth data can give an overview of the distribution of assets and liabilities, which can be highly relevant for financial stability and play a central role in assessing inequality. For this reason, survey data on wealth are of particular importance. Since questions about assets and income are sensitive issues, such surveys often suffer from high item non-response (Riphahn and Serfling, 2005). For example, the Household Finance and Consumption Survey (HFCS) reports for France item non-response rates of nearly 30% for value of saving accounts and largest mortgage on household main residence and almost 80% for current value of household main residence (Household Finance and Consumption Network, 2020a).

Listwise deletion, retaining only records with no items missing, leads to a loss of information, and the remaining units in this dataset are not a good representation of the population, which can lead to biased estimates. Missing values are a problem because the incomplete data do not have the regular (matrix) form needed in almost any statistical method, and therefore handling missing values is necessary. In the literature there are various approaches for dealing with missing data in studies, such as in Rubin (1987) or Longford (2005). Van Buuren (2018) gives an extended overview of approaches to handling and imputing of missing data. Rubin (1976) formulated for the first time the concept of missing data mechanisms by using the indicators of the missing values as random variables and posited a model for them. Methods for missing data are generally based on the assumption that the probability of the missing

data does not depend on the missing values after conditioning on the observed values (MAR). To obtain valid statistical inferences, appropriate assumptions about the mechanism of missing values must be made (Van Buuren, 2018). Two approaches to handling incomplete data are single imputation, where each missing value is imputed once, and multiple imputation (MI), where the missing values are replaced by a small number of plausible values. The advantage of MI is that it reflects the uncertainty of missing data, which is then taken into account in the estimation. There are several surveys of income and wealth data where MI is used, including the Consumer Expenditure Survey, where the income variable is imputed five times (Fisher, 2006), and the HFCS, where also five imputations of the data sets are provided to the user (Household Finance and Consumption Network, 2020a).

Of particular interest may be subpopulations of households, either regionally disaggregated or socio-demographic such as households with particular composition (of ages, gender, labor market status, or educational levels). Various political decisions or global events, such as the financial crisis of 2007/2008 or the COVID-19 pandemic in 2020/2021, may affect these subgroups, usually referred to as areas or domains, to varying degrees. Some of these domains may be represented by very few units in the sample and direct estimators (based only on these subjects) result in a large variance. This issue may be solved by small area estimation (SAE) methods. The model-based estimators used in SAE supplement information from other areas and other data sources. Pfeffermann (2013), Rao and Molina (2015) and Jiang and Rao (2020) give compact overviews and Tzavidis et al. (2018) propose a general framework for the production of small area statistics. SAE methods can be distinguished in unit-level (e.g., Battese et al., 1988) and area-level (Fay and Herriot, 1979) models. Unit-level models have the greater information content, but can only be used when unit-level covariate data are available. In addition, area-level models are often used because they are better suited to account for complex survey designs for point and variance estimates. Therefore, we focus on the Fay-Herriot model in this paper. The Fay-Herriot model can be applied to transformed direct estimators to attain normality of the error terms or to ensure that the resulting estimates are within an appropriate range. Slud and Maiti (2006) and Chandra et al. (2017) study the log-transformed Fay-Herriot model and Sugasawa and Kubokawa (2017) consider a general parametric transformation of the response values. Schmid et al. (2017) use an arcsine transformation to estimate literacy rates of Senegal and Casas-Cordero et al. (2016) to estimate poverty rates of Chile.

In the context of SAE, non-response rates in combination with small sample sizes could have significant influence on the estimates especially with sensitive data such as income and wealth data. The investigation of the integration of the imputation uncertainty into small area estimators has received some attention. Among the publications are, for example, Longford (2004), who uses a multiple hot-deck imputation method in the UK Labour Force Survey to estimate unemployment rates using a small area multivariate shrinkage method. Longford (2005) presents methods for dealing with incomplete data and making inferences using small area estimation methods. An approach to modeling the non-missing at random mechanism in SAE under informative sampling and non-response can be found in Sverchkov and Pfeffermann (2018). Kreutzmann et al. (2019) and Bijlsma et al. (2020) use a Fay-Herriot model with pooled direct estimators after multiple imputation and take into account the additional uncertainty due

to the missing values in the sampling variance. However, both ignore the additional uncertainty in the regression-synthetic part of the model. We extend this approach to address the latter problem in addition to extending the methodology to ratios.

We present an approach in which we combine MI with the transformed Fay-Herriot model. We take the multiply imputed values of the missing values as given by the data provider. To account for the additional uncertainty from imputation, pooled components of the direct estimator are used, as well as pooled components of the regression-synthetic part of the Fay-Herriot model. In particular, the components (direct and regression-synthetic part) are combined for a given transformation in such a way that the resulting MI adjusted model has the known structure of Fay-Herriot models. This approach exploits the existing knowledge about transformations, back-transformations and mean squared error (MSE) approximations of the transformed Fay-Herriot model. We apply the general approach to three special cases relevant to practice and additionally discuss MSE estimators for these special cases:

1. For the general Fay-Herriot model for a mean value, we adapt the Prasad-Rao MSE estimator (Prasad and Rao, 1990) to account for the uncertainty owing to missing values.

2. If the distribution of the target indicator is right-skewed, a log transformation can be used. For this case, we use the adapted Prasad-Rao MSE estimator and apply a back-transformation similar to that presented in Rao and Molina (2015).

3. For the Fay-Herriot model for a ratio with an arcsine transformation, we use insights from Hadam et al. (2020) for the back-transformation of the point estimator, as well as for a parametric bootstrap MSE estimator that can reflect the uncertainty due to the missing values.

The validity of the presented point estimators is demonstrated for the three cases outlined above in a simulation study. It is also shown that the additional uncertainty caused by the missing values is accounted for by the proposed MSE estimators.

The paper is structured as follows. Sections 2.2, 2.3, and 2.4 describe the statistical methodology. In Section 2.2, the transformed Fay-Herriot model is presented, which serves as the basis for the combination with MI. Section 3 describes how the direct and regression-synthetic components of the transformed Fay-Herriot model are combined after MI, which leads to a MI adjusted Fay-Herriot model. In Section 2.4, we consider three common special cases of the model from Section 2.3 and present associated uncertainty measures. The proposed methodology is evaluated in simulation experiments in Section 2.5 and then applied to HFCS data in Section 2.6. Section 2.7 summarizes the main findings, discusses limitations of the approach and outlines further research potential.

2.2 Transformed Fay-Herriot model

In the following the transformed Fay-Herriot model is introduced, where the transformation is described by a known function h. Let N be the size of a finite population which is partitioned into $d = 1, ..., D$ domains and n the sample size with $i = 1, ..., n_d$ units per domain so that $n = \sum_{d=1}^{D} n_d$. The Fay-Herriot model involves in the first stage a sampling model in which it

is supposed that the direct estimator consists of the true domain-specific population indicator θ_d and a sampling error e_d:

$$\hat{\theta}_d^{Dir} = \theta_d + e_d, \quad e_d \overset{ind}{\sim} N\left(0, \sigma_{e_d}^2\right).$$

It is assumed that the sampling errors e_d are independently normally distributed with known variance $\sigma_{e_d}^2$. Although the sampling variances $\sigma_{e_d}^2$ are assumed to be known, in practice they are estimated by unit-level data (Rivest and Vandal (2002), Wang and Fuller (2003), You and Chapman (2006)). Another unit-level approach to address the problem of unknown sampling variances is proposed by Maiti et al. (2014) and Sugasawa et al. (2017) by shrinking and simultaneous modeling of small area means and variances. When the indicator of interest is a mean value, a domain specific direct estimator is the weighted average of the sampled values:

$$\hat{\theta}_d^{Dir} = \frac{\sum_{i=1}^{n_d} w_{id} y_{id}}{\sum_{i=1}^{n_d} w_{id}}.$$

The incorporation of sampling weights w_{id} makes the point estimator design unbiased. Note that the population and the outcomes y_{id} are assumed to be fixed, and the sampling mechanism is the only source of uncertainty. The sampling weights reflect a complex design in the estimation of the associated variance. The second stage of the Fay-Herriot model is a linking model, which links covariate information to the population indicator. x_d is a $p \times 1$ vector with area-level population covariates and β is the corresponding $p \times 1$ vector with regression coefficients. v_d are normally distributed domain specific random effects:

$$\theta_d = x_d^T \beta + v_d, \quad v_d \overset{iid}{\sim} N\left(0, \sigma_v^2\right). \tag{2.1}$$

Combining the sampling and the linking model results in:

$$\hat{\theta}_d^{Dir} = x_d^T \beta + v_d + e_d, \quad v_d \overset{iid}{\sim} N\left(0, \sigma_v^2\right), \quad e_d \overset{ind}{\sim} N\left(0, \sigma_{e_d}^2\right). \tag{2.2}$$

If a smooth and monotone transformation function h is applied to the direct estimator, $\hat{\theta}_d^{Dir}$ is replaced by $\hat{\theta}_d^{Dir*} := h\left(\hat{\theta}_d^{Dir}\right)$ in Equation (2.2) and we want to predict $h^{-1}(\theta_d)$. The transformed Fay-Herriot model is then defined, for example, as in Sugasawa and Kubokawa (2017):

$$h\left(\hat{\theta}_d^{Dir}\right) = x_d^T \beta + v_d + e_d, \quad v_d \overset{iid}{\sim} N\left(0, \sigma_v^2\right), \quad e_d \overset{ind}{\sim} N\left(0, \sigma_{e_d}^{2*}\right). \tag{2.3}$$

In the following, * always refers to the transformed scale of the direct estimator, its variance and the Fay-Herriot estimator presented at the end of this section. The model parameters, the model variance σ_v^2 and the regression coefficients β are not known and must be estimated. There are various methods to obtain estimates of σ_v^2, for example, restricted maximum likelihood (REML), maximum likelihood (ML) and the FH method-of-moments. More details on the estimation methods of the model variance can be found in Chapter 6 in Rao and Molina (2015). A drawback of ML is that it does not account for the loss in degrees of freedom arising from the estimation of the regression coefficients β (Rao and Molina, 2015). Therefore, we use in

this paper the REML method. The regression coefficients β and the random effects v_d are estimated by:

$$\hat{\beta} = \hat{\beta}\left(\hat{\sigma}_v^2\right) = \left(\sum_{d=1}^{D} \frac{x_d x_d^T}{\sigma_{e_d}^{2*} + \hat{\sigma}_v^2}\right)^{-1} \left(\sum_{d=1}^{D} \frac{x_d \hat{\theta}_d^{Dir*}}{\sigma_{e_d}^{2*} + \hat{\sigma}_v^2}\right), \tag{2.4}$$

$$\hat{v}_d = \frac{\hat{\sigma}_v^2}{\sigma_{e_d}^{2*} + \hat{\sigma}_v^2}\left(\hat{\theta}_d^{Dir*} - x_d^T \hat{\beta}\right). \tag{2.5}$$

Plugging those predictors into Equation (2.1) leads to the empirical best linear unbiased predictor (EBLUP), i.e. the transformed Fay-Herriot estimator:

$$\hat{\theta}_d^{FH*} = x_d^T \hat{\beta} + \hat{v}_d. \tag{2.6}$$

This estimator can be expressed as a convex combination of the direct estimator and the regression-synthetic component, resulting in an optimal combination of the two components. If the variance of the direct estimator is large, more weight is given to the synthetic component, and vice versa:

$$\hat{\theta}_d^{FH*} = \hat{\gamma}_d \hat{\theta}_d^{Dir*} + (1 - \hat{\gamma}_d)\, x_d^T \hat{\beta} \quad \text{with} \quad \hat{\gamma}_d = \frac{\hat{\sigma}_v^2}{\sigma_{e_d}^{2*} + \hat{\sigma}_v^2}. \tag{2.7}$$

At this point $\hat{\theta}_d^{FH*}$ is still on the transformed scale and has to be transformed to the original scale to obtain $\hat{\theta}_d^{FH}$.

2.3 Combining transformed Fay-Herriot models after multiple imputation

An often applied technique to handle missing values is MI, where the missing values are replaced by several plausible values. To obtain these values, an imputation model is required. It is not sufficient to generate only one imputation, since the imputation is treated as if it were true, and the uncertainties arising from the non-response are ignored. On the contrary, a large number of imputations is usually not necessary, and M between 5 and 20 is sufficient, but it may be advantageous to choose a higher value (20 - 100) if the non-response is high and there is a large uncertainty about the estimand (Van Buuren, 2018). The procedure for MI involves two steps: the imputation step and the analysis step. In the former, the imputer, usually the data provider, generates the M replicate completions of the survey data using a suitable imputation model and provides them to the analyst. In the second step, the analyst applies a statistical model suitable for the complete data separately to each imputed data set. The focus of this paper is on the latter. If θ is the indicator of interest and $\hat{\theta}$ its estimator, the analysis model is calculated with each imputed data set, so we obtain $\hat{\theta}_m$ and $\widehat{\text{Var}}\left(\hat{\theta}_m\right)$ for $m = 1, ..., M$. The results are then combined with the application of pooling rules developed by Rubin (1987) for point estimates and their variances, which include the additional variability and uncertainty induced by the missing data. Rubin's rules (RR) are defined as follows. The pooled estimator

of θ is the mean value of the M estimators:

$$\hat{\theta}^{RR} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m. \tag{2.8}$$

The variance of the pooled estimator $\hat{\sigma}^{2^{RR}}$ is composed by the mean value of the individual variances of each estimator (within-variance) and the variance between the M estimates (between-variance) with an correction due to the finite sample size:

$$\hat{\sigma}^{2^{RR}} = \widehat{\text{Var}}\left(\hat{\theta}^{RR}\right) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\text{Var}}\left(\hat{\theta}_m\right) + \frac{M+1}{M} \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \hat{\theta}^{RR}\right)^2. \tag{2.9}$$

In the next sections, we describe how the combining rules are applied to the components of the transformed Fay-Herriot model from Section 2.2.

2.3.1 Component pooling

With the M multiply imputed sampling values $y_{id,m}$ of each unit $i = 1, ..., n_d$ and domains $d = 1, ..., D$, the transformed direct estimators $\hat{\theta}^{Dir*}_{d,m} = h\left(\hat{\theta}^{Dir}_{d,m}\right)$ of the target indicator and their corresponding sample variances $\sigma^{2*}_{e_{d,m}}$ are calculated for each domain $d = 1, ..., D$ and $m = 1, ..., M$. Rubin's rules are based on asymptotic theory, and the resulting combined estimate is more accurate if the distribution of the indicator of interest is better approximated by the normal distribution (Rubin, 1987). Van Buuren (2018) states that to promote approximate normality, target indicators can be transformed, then pooled and back-transformed. Therefore, the M direct estimators $\hat{\theta}^{Dir*}_{d,m}$ and their variances $\sigma^{2*}_{e_{d,m}}$ are pooled on the transformed scale and substituted in Equations (2.8) and (2.9). Kreutzmann et al. (2019) present a Fay-Herriot estimator which uses pooled direct components on the original scale, which are substituted in the (log transformed) Fay-Herriot model. We extend this approach and transform the direct components of each imputed data set to estimate the regression-synthetic components. This allows the uncertainty of the missing values to be included not only in the direct components, but also in those of the linking model. The model components of the linking model are estimated for each imputed data set. The estimated variances of the random effects $\hat{v}_{d,m}$ are combined according to Rubin's rule:

$$W = \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}^2_{v_m} \quad \text{and} \quad B_d = \frac{M+1}{M} \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{v}_{d,m} - \frac{1}{M} \sum_{m=1}^{M} \hat{v}_{d,m} \right)^2. \tag{2.10}$$

The mean squared distance of the random effects of the domains of the M imputed data sets and the pooled random effects per domain is different between the areas. In order to guarantee that the random effects have a common variance, further pooling has to be applied. Therefore, the mean value of the between variance is taken. Together with Equation (2.10) this leads to the pooled model variance:

$$\hat{\sigma}^{2^{RR}}_v = W + \frac{1}{D} \sum_{d=1}^{D} B_d. \tag{2.11}$$

The pooled model variance $\hat{\sigma}_v^{2^{RR}}$ and the pooled direct components are now used to obtain MI adjusted estimates of the regression coefficients and random effects. $\hat{\sigma}_v^{2^{RR}}$, $\hat{\sigma}_{e_d}^{2^{RR*}}$ and $\hat{\theta}_d^{Dir.RR*}$ are inserted into Equation (2.4) to obtain the MI adjusted regression coefficients $\hat{\beta}$ and then together into Equation (2.5) to obtain the MI adjusted random effects $\hat{v}_d$.

2.3.2 MI adjusted Fay-Herriot model

The pooled direct components together with the pooled and MI adjusted regression-synthetic parts of the model lead to the MI adjusted Fay-Herriot model, which preserves the structure of the transformed Fay-Herriot model. The area-level population auxiliary information x_d, obtained from external sources, such as the census, is fixed and complete as in Equation (2.1). The model can be written analogously to Equation (2.3) with pooled direct components and the pooled model variance. Using the estimators of unknown model parameters as elaborated in Section 2.3.1 leads to the proposed FH.MI estimator $\hat{\theta}_d^{FH.MI*}$, which can be written analogously to Equation (2.7) with $\hat{\theta}_d^{Dir.RR*}$, $\hat{\sigma}_{e_d}^{2^{RR*}}$ and $\hat{\sigma}_v^{2^{RR}}$ plugged in:

$$\hat{\theta}_d^{FH.MI*} = \hat{\gamma}_d \hat{\theta}_d^{Dir.RR*} + (1 - \hat{\gamma}_d)\, x_d^T \hat{\beta} \quad \text{with} \quad \hat{\gamma}_d = \frac{\hat{\sigma}_v^{2^{RR}}}{\hat{\sigma}_{e_d}^{2^{RR*}} + \hat{\sigma}_v^{2^{RR}}}. \tag{2.12}$$

The presented $\hat{\theta}_d^{FH.MI*}$ estimator preserves the representations of the Fay-Herriot estimator. As $\hat{\theta}_d^{FH.MI*}$ is on the transformed scale, a suitable back transformation depending on h has to be applied to obtain $\hat{\theta}_d^{FH.MI}$.

Small area estimators with multiply imputed data can be derived in two ways: 1. Fit the Fay-Herriot model to each of the M imputed data sets and combine the Fay-Herriot estimators with Rubin's rule. 2. Estimate the direct and the regression synthetic components M times and combine them using Rubin's rules as described in Section 2.3.1 and then estimate the shrinkage estimator in Equation (2.12). The advantage of the first approach is that it is simple. However, it loses the structure of the Fay-Herriot model and the representation of the estimator as a weighted combination of the direct and regression synthetic components. In addition, it is unclear how the uncertainty of the M Fay-Herriot estimators is combined, since Rubin's rule is commonly used for variances and it is unclear how this rule can be applied to the MSE. The advantage of the second (the proposed) approach and the resulting FH.MI estimator is that the model structure of the Fay-Herriot model is preserved, the interpretability of the components is maintained, and the existing knowledge about MSE estimators is directly transferable and extensible. The estimator of the first approach is used as a benchmark in the model-based simulation study in Section 2.5 and is denoted by FH.RR.

2.4 MI adjusted Fay-Herriot estimators with uncertainty measures

In the following sections, we focus on three special cases of the transformed MI adjusted Fay-Herriot estimator (2.12). For each case we specify the FH.MI point estimator and an associated MSE estimator.

2.4.1 Estimator for a mean

The (population) mean of a quantity of interest for domain d is estimated by the weighted sample average per imputed data set m:

$$\hat{\theta}_{d,m}^{Dir} = \frac{\sum_{i=1}^{n_d} w_{id} y_{id,m}}{\sum_{i=1}^{n_d} w_{id}} \quad \text{for} \quad d = 1, .., D \quad \text{and} \quad m = 1, ..., M. \tag{2.13}$$

If no transformation is required for the direct estimator, $\hat{\theta}_d^{FH.MI*}$ is on the original scale such that $\hat{\theta}_d^{FH.MI} = \hat{\theta}_d^{FH.MI*}$. With the pooled and MI adjusted estimators presented in Section 2.3, the FH.MI estimator $\hat{\theta}_d^{FH.MI}$ can be calculated according to Equation (2.12). As a measure of uncertainty which captures the additional uncertainty due to multiple imputation, we adapt the MSE estimator of Prasad and Rao (1990) in the following. The second-order approximation of the MSE of $\hat{\theta}_d^{FH}$ is given by:

$$\text{MSE}\left(\hat{\theta}_d^{FH}\right) \approx g_{1d}\left(\sigma_v^2\right) + g_{2d}\left(\sigma_v^2\right) + g_{3d}\left(\sigma_v^2\right).$$

The first component g_{1d} is based on the prediction of the random effects and g_{2d} reflects the variability arising from the estimation of the regression coefficients. g_{1d} and g_{2d} are independent of the estimation method of the model variance σ_v^2, whereas, g_{3d} reflects the uncertainty caused by the estimation of σ_v^2 and depends on the estimation method through its asymptotic variance $\bar{V}\left(\hat{\sigma}_v^2\right)$ (as $D \to \infty$) (see e.g., Rao and Molina (2015)). According to Prasad and Rao (1990) a second-order unbiased estimator of MSE $\left(\hat{\theta}_d^{FH}\right)$ is:

$$\widehat{\text{MSE}}\left(\hat{\theta}_d^{FH}\right) = g_{1d}\left(\hat{\sigma}_v^2\right) + g_{2d}\left(\hat{\sigma}_v^2\right) + 2g_{3d}\left(\hat{\sigma}_v^2\right).$$

The components of the Prasad-Rao estimator using REML are defined as follows:

$$g_{1d}\left(\hat{\sigma}_v^2\right) = \hat{\gamma}_d^2 \sigma_{e_d}^2, \tag{2.14}$$

$$g_{2d}\left(\hat{\sigma}_v^2\right) = \left(1 - \hat{\gamma}_d\right)^2 x_d^T \left\{\sum_{d=1}^{D} \frac{x_d x_d^T}{\sigma_{e_d}^2 + \hat{\sigma}_v^2}\right\}^{-1} x_d, \tag{2.15}$$

$$g_{3d}\left(\hat{\sigma}_v^2\right) = \left(\sigma_{e_d}^2\right)^2 \left(\sigma_{e_d}^2 + \hat{\sigma}_v^2\right)^{-3} \bar{V}\left(\hat{\sigma}_v^2\right), \tag{2.16}$$

$$\bar{V}\left(\hat{\sigma}_v^2\right) = 2\left\{\sum_{d=1}^{D} \frac{1}{\left(\sigma_{e_d}^2 + \hat{\sigma}_v^2\right)^2}\right\}^{-1}.$$

In the same way as in Section 2.3.1, where we obtain M estimates of the model variance, i.e., $\hat{\sigma}_{v_m}^2$ for $m = 1, ..., M$, we obtain M corresponding asymptotic ($D \to \infty$) variances $\bar{V}_m\left(\hat{\sigma}_{v_m}^2\right)$ for $m = 1, ..., M$. To adjust the MSE estimator for this additional uncertainty, the asymptotic

variances are pooled with Rubin's rule for variances (2.9):

$$\bar{V}^{RR}\left(\hat{\sigma}_v^{2^{RR}}\right) = \frac{1}{M}\sum_{m=1}^{M}\bar{V}_m\left(\hat{\sigma}_{v_m}^2\right) + \frac{M+1}{M}\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\sigma}_{v_m}^2 - \hat{\sigma}_v^{2^{RR}}\right)^2$$
$$\text{with} \quad \bar{V}\left(\hat{\sigma}_{v_m}^2\right) = 2\left\{\sum_{d=1}^{D}\frac{1}{\left(\sigma_{e_{d,m}}^2 + \hat{\sigma}_{v_m}^2\right)^2}\right\}^{-1} \quad \text{for} \quad m = 1, ..., M. \tag{2.17}$$

Using $\hat{\sigma}_v^{2^{RR}}$ and $\sigma_{e_d}^{2^{RR}}$ in (2.14), (2.15), and (2.16) together with the pooled asymptotic variance (2.17) takes into account the uncertainty about the missing values. Note that instead of plugging the pooled variance terms into the asymptotic variance formula, the pooled asymptotic variance $\bar{V}^{RR}\left(\hat{\sigma}_v^{2^{RR}}\right)$ is used, introducing an additional term into the estimator due to the between-variation. This leads to the proposed MSE estimator for $\hat{\theta}_d^{FH.MI}$, which captures the uncertainty due to missing values:

$$\begin{aligned}\widehat{\text{MSE}}\left(\hat{\theta}_d^{FH.MI}\right) &= g_{1d}\left(\hat{\sigma}_v^{2^{RR}}\right) + g_{2d}\left(\hat{\sigma}_v^{2^{RR}}\right) \\ &\quad + 2\left(\sigma_{e_d}^{2^{RR}}\right)^2\left(\sigma_{e_d}^{2^{RR}} + \hat{\sigma}_v^{2^{RR}}\right)^{-3}\bar{V}^{RR}\left(\hat{\sigma}_v^{2^{RR}}\right).\end{aligned} \tag{2.18}$$

2.4.2 Estimator for a log mean

Domain specific mean values of income and wealth data are often skewed to the right, or the relationship with the auxiliary information may be non-linear. In such a case, the linear Fay-Herriot model (Section 2.4.1) may be more appropriate for the log-transformed direct estimator. Using the direct estimator from Equation (2.13) and $h\colon z \mapsto log\,(z)$ the direct components of the model for the M imputed data sets are:

$$\hat{\theta}_{d,m}^{Dir*} = log\left(\hat{\theta}_{d,m}^{Dir}\right) \quad \text{with variances} \quad \sigma_{e_{d,m}}^{2*} \approx \left(\hat{\theta}_{d,m}^{Dir}\right)^{-2}\sigma_{e_{d,m}}^2$$
$$\text{for} \quad d = 1, .., D, \quad m = 1, ..., M.$$

Using a Taylor expansion for moments, the sample variance, i.e., the variance of the direct estimator, can be moved to the logarithmic scale. Although this is an approximation for large samples, it is used in SAE as in Neves et al. (2013). Council (2000) use the same approximation with a minor modification based on the properties of the log-normal distribution, while noting that the results do not differ considerably. Calculating the direct and the regression-synthetic components as described in Section 2.3.1 with $h\colon z \mapsto log\,(z)$ and together with Equation (2.12) leads to the Fay-Herriot-MI estimator $\hat{\theta}_d^{FH.MI*}$, which is still on the log-scale. The estimates can be transformed back to the original scale by several methods. Slud and Maiti (2006) present a bias-correction under a log-transformed Fay-Herriot model and propose a corresponding estimator for the MSE. Chandra et al. (2017) extend this estimator by an additional bias correction that accounts for the sampling variation of the estimator. These methods can be applied only to observed/sampled areas. We apply a method that is suitable even for domains/areas with no observations. To obtain the point estimator on the original scale, proper-

ties of the log-normal distribution are used and the back-transformation for the MSE estimator is based on a Taylor expansion similar to that presented in Rao and Molina (2015). A short derivation can be found in the Appendix. The back-transformation is defined as follows :

$$\hat{\theta}_d^{FH.MI} = exp\left\{\hat{\theta}_d^{FH.MI*} + 0.5\widehat{\text{MSE}}\left(\hat{\theta}_d^{FH.MI*}\right)\right\},$$
$$\widehat{\text{MSE}}\left(\hat{\theta}_d^{FH.MI}\right) = exp\left\{\hat{\theta}_d^{FH.MI*} + 0.5\widehat{\text{MSE}}\left(\hat{\theta}_d^{FH.MI*}\right)\right\}^2 \widehat{\text{MSE}}\left(\hat{\theta}_d^{FH.MI*}\right).$$

$\widehat{\text{MSE}}\left(\hat{\theta}_d^{FH.MI*}\right)$ denotes at this point the adapted Prasad-Rao MSE estimator defined in Equation (2.18).

2.4.3 Estimator for an arcsine ratio

The Fay-Herriot model is widely used for estimating poverty or literacy rates with high regional resolution. In order to guarantee that the estimated rates are between 0 and 1 suitable transformations are frequently used. The arcsine transformation $h\colon z \mapsto sin^{-1}(\sqrt{z})$, of which the inverse maps its values to $[0, 1]$, is commonly used. Schmid et al. (2017) compared in a design-based simulation the arcsine transformation with an estimator based on a normal-logistic distribution. Both estimators provided very similar results regarding bias and root mean squared error (RMSE). We concentrate on the arcsine transformation because, unlike the logit, it is well defined even at zero and unity. The arcsine transformation is applied to the direct ratio estimators of the M imputed data sets:

$$\hat{\theta}_{d,m}^{Dir*} = sin^{-1}\left(\sqrt{\hat{\theta}_{d,m}^{Dir}}\right) \quad \text{with variances} \quad \sigma_{e_{d,m}}^{2*} = \sigma_{e_d}^{2*} = \frac{1}{4\tilde{n}_d} \quad \text{for} \quad m = 1, .., M.$$

The effective sample size of domain d is denoted by $\tilde{n}_d$, which takes into account the sampling design effect (Jiang et al., 2001). The approximation of the sampling error variance on the transformed scale is based on a Taylor expansion for moments like in Jiang et al. (2001). The combined point estimator $\hat{\theta}_d^{Dir.RR*}$ and its variance $\hat{\sigma}_{e_d}^{2^{RR*}}$ are calculated by applying Rubin's rules presented in Equations (2.8) and (2.9). The components of the regression-synthetic part of the model are calculated as described in Section 2.3.1 with the pooled direct components on the transformed scale. Afterwards $\hat{\theta}_d^{FH.MI*}$ can be calculated as in Equation (2.12). The resulting estimator $\hat{\theta}_d^{FH.MI*}$ is on a $sin^{-1}\left(\sqrt{}\right)$-scale and needs to be transferred to the original scale. A naive back-transformation is the inverse h^{-1}, which introduces a bias for non-linear h. For this reason, for common transformations bias-corrected back-transformations are proposed, such as in Hadam et al. (2020) for the arcsine transformation which is a special case of Sugasawa and Kubokawa (2017), who present an asymptotically unbiased back-transformation for a general parametric transformation. We apply the bias-corrected back-transformation following Hadam et al. (2020), using the normal distribution of the transformed estimator and the expected value

(E) of a transformed variable:

$$\hat{\theta}_d^{FH.MI} = \mathrm{E}\left[sin^2\left(\hat{\theta}_d^{FH.MI*}\right)\right] = \int_{-\infty}^{\infty} sin^2(t)\, f_{\hat{\theta}_d^{FH.MI*}}(t)\, dt$$
$$= \int_{-\infty}^{\infty} sin^2(t) \frac{1}{\sqrt{2\pi \frac{\hat{\sigma}_v^{2RR}\sigma_{e_d}^{2RR*}}{\hat{\sigma}_v^{2RR}+\sigma_{e_d}^{2RR*}}}} exp\left\{-\frac{\left(t-\hat{\theta}_d^{FH.MI*}\right)^2}{2\frac{\hat{\sigma}_v^{2RR}\sigma_{e_d}^{2RR*}}{\hat{\sigma}_v^{2RR}+\sigma_{e_d}^{2RR*}}}\right\} dt. \tag{2.19}$$

The integral in Equation (2.19) must be solved by numerical integration methods. The MSE of $\hat{\theta}_d^{FH.MI}$ is approximated with a parametric bootstrap procedure analogue to Hadam et al. (2020) based on Gonzalez-Manteiga et al. (2008b). The bootstrap procedure comprises the following steps:

1. Estimate the regression-synthetic components $\hat{\beta}$ and $\hat{\sigma}_v^{2^{RR}}$ analogously to Section 2.3.1 using the pooled direct components $\hat{\theta}_d^{Dir.RR*}$ and $\hat{\sigma}_{e_d}^{2^{RR*}}$ on the arcsine scale.

2. For $b = 1, ..., B$

   (a) Generate sampling errors $e_d^{(b)} \overset{ind}{\sim} N\left(0, \hat{\sigma}_{e_d}^{2^{RR*}}\right)$ and random effects $v_d^{(b)} \overset{iid}{\sim} N\left(0, \hat{\sigma}_v^{2^{RR}}\right)$.

   (b) Simulate a bootstrap sample $\hat{\theta}_d^{Dir*(b)} = x_d^T\hat{\beta} + v_d^{(b)} + e_d^{(b)}$.

   (c) Calculate the true bootstrap population indicator $\theta_d^{*(b)} = x_d^T\hat{\beta} + v_d^{(b)}$ on the transformed scale and back-transform with $\theta_d^{(b)} = sin^2\left(\theta_d^{*(b)}\right)$.

   (d) Calculate the bootstrap estimator of the model variance $\hat{\sigma}_v^{2(b)}$ using $\hat{\theta}_d^{Dir*(b)}$ and $\hat{\sigma}_{e_d}^{2^{RR*}}$.

   (e) Using $\hat{\sigma}_v^{2(b)}$ and $\hat{\theta}_d^{Dir*(b)}$, calculate bootstrap estimators of the regression coefficients $\hat{\beta}^{(b)}$ and estimate the random effects $\hat{v}_d^{(b)}$.

   (f) Determine the bootstrap estimator $\hat{\theta}_d^{FH.MI*(b)}$ with Equation (2.12) by using the estimates from the step before and back-transform to the original scale applying (2.19) to obtain $\hat{\theta}_d^{FH.MI(b)}$.

3. Estimate the MSE:

$$\widehat{\mathrm{MSE}}(\hat{\theta}_d^{FH.MI}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_d^{FH.MI(b)} - \theta_d^{(b)}\right)^2.$$

The pooled sampling and model variances, which account for the additional uncertainty about the missing values, are used in the initialization of the bootstrap method. Hence, the extra uncertainty induced by the missing data is accounted for by the bootstrap MSE estimator.

2.5 Simulation study

In this section, we investigate the behaviour of the estimators proposed in Sections 2.3 and 2.4 by simulation studies with suitable data models. The population is repeatedly generated

according to an underlying model. With each simulation run, a sample is taken from the generated population, to which the methods are then applied. We evaluate the performance in terms of bias and RMSE of the proposed point estimators and the inflation of RMSE arising from MI.

2.5.1 Data generation

The simulation setup and data models are chosen to be consistent with those of Kreutzmann et al. (2019). For the simulations, finite populations of size $N = 60,000$ with $D = 100$ domains are generated so that in each domain the population size N_d is between 200 and 1000 for $d = 1, ..., D$. The samples were drawn via stratified random sampling, where the strata represent the domains. To have rather small and large domains in the samples, sample sizes n_d lie within a range of 8 and 145, so that the total sample size is $n = 5961$. To apply the transformations discussed in the special cases in Section 2.4, appropriate data models are chosen. In the standard case, a normal data model is used, where no transformation to the direct estimator of a mean value is necessary. Right-skewed log-normal data is generated when investigating the proposed method with a log transformation like in Section 2.4.2. In many applications, the indicator of interest is a ratio. In order to construct a ratio that is used in real data applications, a wealth ratio is calculated. In publications of the Federal Statistical Office (see e.g., Destatis (2018)) it is derived by taking the percentage of households with a household income above the 200% median household income. As data model for the ratio the log-scale data is also used. The unit-level data models and scenarios are described in detail in Table 2.1. The shapes of the distribution for one selected population can be found in Figure B.1 in the Appendix. With a sample at the unit-level, the missing data is generated. As mentioned in

Table 2.1: Overview of unit-level data models in model-based simulation, $i = 1, ..., N, \quad d = 1, ..., D$.

Setting	y_{id}	x_{id}	μ_d	v_d	e_{id}
mean	$250000 - 400x_{id} + v_d + e_{id}$	$N(\mu_d; 150^2)$	$U[-150, 150]$	$N(0, 25000^2)$	$N(0, 50000^2)$
log mean	$exp(15 - x_{id} + v_d + e_{id})$	$N(\mu_d; 1)$	$U[3, 5]$	$N(0, 0.4^2)$	$N(0, 0.6^2)$
ratio	$exp(15 - x_{id} + v_d + e_{id})$	$N(\mu_d; 1)$	$U[3, 5]$	$N(0, 0.4^2)$	$N(0, 0.6^2)$

Section 2.1, MAR is often plausible and assumed in most programs for handling missing data. Therefore, in the simulation, missing values are generated using the fully observed additional variable x, from the data models in Table 2.1. The MAR mechanism is implemented as follows:

$$y_{id} = \begin{cases} y_{\text{missing}}, & x_{id} \leq x_q \\ y_{id}, & \text{otherwise.} \end{cases} \tag{2.20}$$

x_q is the q-quantile of the auxiliary information x from the sample. This results in a non-response rate of $q \cdot 100\%$ by definition of the q-quantile. For the selected data models, the implemented MAR mechanism leads to missing values in the upper ends of the distribution. When it comes to sensible data as wealth related data, item non-response rates can be very high. For example, the Household Finance and Consumption Network (HFCN) reports for 2017 (Household Finance and Consumption Network, 2020a) non-response rates for the value of savings account between 18% in Belgium and 64% in Finland. Therefore, it is reasonable

to investigate the proposed methods under varying $q \in \{0.1, 0.3, 0.5\}$ to obtain non-response rates of 10%, 30% and 50%. A two-level normal model is used as an imputation model for the missing y_{id} values, which is implemented in the R-package `mice` (Van Buuren and Groothius-Oudshoorn, 2011). The x serve as covariate information and v_d as area-specific random effects, so that the clustering is incorporated in the imputation model. According to Van Buuren (2018), between five and 20 imputed values are often sufficient for each missing observation. The HFCN delivers five imputed values per missing observation, hence in the simulation we set $M = 5$. In the log-scale setting the data was log transformed prior to the imputation to achieve normality and back transformed with the inverse afterwards. After imputation, the data is still on a unit-level and has to be aggregated on an area-level according to the indicator of interest of the setting. Then the appropriate FH.MI estimators given in Section 2.3 with the special cases in Section 2.4 are calculated. Table 2.2 provides an overview showing for each setting the direct estimator, the transformation used, and the section of the corresponding FH.MI model for the special case. In Table 2.2, I denotes an indicator function that is 1 if the condition is true and 0 otherwise; $\tilde{Y}$ denotes the population median of y.

Table 2.2: Overview of settings.

Setting	$\hat{\theta}_d^{Dir}$	$h\left(\hat{\theta}_d^{Dir}\right)$	FH.MI model
mean	$\frac{1}{n_d}\sum_{i=1}^{n_d} y_{id}$	$\hat{\theta}_d^{Dir}$	2.4.1
log mean	$\frac{1}{n_d}\sum_{i=1}^{n_d} y_{id}$	$log\left(\hat{\theta}_d^{Dir}\right)$	2.4.2
ratio	$\frac{1}{n_d}\sum_{i=1}^{n_d} I\left(y_{id} > 2 \cdot \tilde{Y}\right)$	$sin^{-1}\left(\sqrt{\hat{\theta}_d^{Dir}}\right)$	2.4.3

Each setting, including the generation of the population according to the data model, the sampling, the missing data generating process, the multiple imputation and the application of the MI adjusted FH estimators is repeated $R = 500$ times. The steps of the simulation can be summarized as follows: We generate the population according to a data model in Table 2.1. Next a stratified random sample is selected. Then missing values are generated according to Equation (2.20) and imputed to create M copies of the data. Using the M data sets the direct estimators are calculated according to Table 2.2 and x_{id} are aggregated to a domain level by taking the mean per domain. Afterwards the indicator of interest and its MSE are estimated by applying the methods described in Sections 2.3 and 2.4.

2.5.2 Performance of point estimators

In the simulation we assess the performance of six point estimators in the *mean* and *log mean* setting and five in the *ratio* setting. For each setting direct, (Direct) and Fay-Herriot (FH) estimators are calculated before deletion on the aggregated sample, that is, the steps of deleting and imputing are omitted. In the case of the FH estimator, the transformation corresponding to the setting is applied so that the Fay-Herriot estimator introduced in Section 2.2 is calculated. The FH estimator before deletion serves as the gold standard in this simulation. In addition, we compare the performance of the proposed FH.MI estimators with the pooled Fay-Herriot estimator (FH.RR) mentioned in Section 2.3 and with the estimator proposed by Kreutzmann et al.

(2019) denoted by FH.DirectRR. They consider the estimator under a normal and log-normal setting for a mean value, and so we also examine this estimator only under these settings. Furthermore, with Rubin's rule combined direct estimators (Direct.RR) are calculated to show the efficiency gain of the Fay-Herriot estimators with good covariate information after MI. All estimators are implemented in the statistical programming language R (R Core Team, 2022) and for the standard area-level models and its components the package **emdi** (Kreutzmann et al., 2019) was used. The code can be obtained from the authors on request. To evaluate and compare the performance of the estimators, the following quality measures are calculated using the R Monte-Carlo replications. $\hat{\theta}_{d_r}$ denotes the estimator of the target indicator in domain d and replication r, θ_{d_r} is the true value of the indicator:

$$\begin{aligned} \text{Bias}(\hat{\theta}_d) &= \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{d_r} - \theta_{d_r}\right), \text{ rel. Bias}(\hat{\theta}_d) = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\theta}_{d_r} - \theta_{d_r}}{\theta_{d_r}}\right), \\ \text{RMSE}(\hat{\theta}_d) &= \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{d_r} - \theta_{d_r}\right)^2}, \text{ RRMSE}(\hat{\theta}_d) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\theta}_{d_r} - \theta_{d_r}}{\theta_{d_r}}\right)^2}. \end{aligned} \tag{2.21}$$

We want to evaluate the performance of the introduced methodology in terms of bias and RMSE. For the *mean* and *log mean* setting we consider the relative bias and the RRMSE. For the *ratio* setting the bias and RMSE are taken into account since the indicator itself is already on a relative scale. The median and mean values over domains of the bias and RMSE values for different non-response rates are presented in Table 2.3. The direct estimators (Direct.RR) remain unbiased after multiple imputation in the *mean* and *ratio* setting as before deletion (Direct) and almost unbiased in the *log mean* setting. The small bias could be introduced by the inverse back-transformation after applying the imputation model. Compared to the combined direct estimators (Direct.RR) and the model-based estimators before deletion (FH), the model-based estimators FH.MI, FH.RR and FH.DirectRR remain also unbiased in the *mean* and *ratio* setting and the results of the model-based estimators are comparable. Only in the *log mean* setting does the FH.MI estimator, like the other two model-based estimators, suffer from a small bias that increases slightly with higher non-response rates. Again this bias could be due to the inverse back-transformation in the imputation process. In terms of efficiency, we see that the RRMSE/RMSE are the smallest before deletion and increase with higher non-response rates for each estimator in each setting, reflecting the additional uncertainty about missing values. Within each setting and non-response rate the order of the RRMSE/RMSE is as expected: the RRMSE/RMSE of the direct estimators is always higher than that of the proposed FH.MI estimator, which shows that the introduced methodology behaves the same way as in cases without missing values (i.e., before deletion). The RRMSE/RMSE of the FH.MI and the FH.RR are almost identical, which indicates that the proposed methodology leads to reasonable results and is similar to the more straightforward approach of combining the Fay-Herriot estimators. The proposed FH.MI estimator is at least as efficient as the FH.DirectRR estimator. In the *log mean* setting, the *superefficiency* of imputation, when more information is used than in the analysis model (Rubin, 1996), can be observed. At a non-response rate of 10%, Direct.RR is slightly more efficient than the direct estimator before deletion (Direct). All summed up, the results

confirm our expectations. The presented FH.MI estimators lead to plausible results regarding bias and efficiency in the investigated settings, in which the imputation models follow the data structure of the generated population and thus fit the data.

2.5.3 Performance of uncertainty measures

We now move on to the performance of the three proposed MSE estimators of the FH.MI estimator, each corresponding to one setting. In the case of the *mean* and *log mean* setting, we evaluate the adapted analytical Prasad-Rao estimator as described in Sections 2.4.1 and 2.4.2 with a back-transformation when the log transformation is used. In the *ratio* setting the parametric bootstrap estimator from Section 2.4.3 with $B = 500$ replications is evaluated. Performance is evaluated by looking at the relative bias of the MSE estimator defined as followed:

$$\text{RBRMSE}(\hat{\theta}_d) = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R} \widehat{\text{MSE}}_{d_r}} - \text{RMSE}(\hat{\theta}_d)}{\text{RMSE}(\hat{\theta}_d)}.$$

Table 2.4 shows the median and mean values over the domains of the RBRMSE. We see a slight underestimation in the *mean* setting with an increasing effect at higher non-response rates. On the other hand, in the *log mean* setting the true RMSE is slightly overestimated at a lower non-response rate of 10% and minimally underestimated at a higher non-response rate of 50%. Nevertheless, the values are all close to zero. In the *ratio* setting, the bias of the bootstrap RMSE estimator is close to zero at 10% non-response rate. At 30% and 50% it increases and reaches almost identical values, but still at a tolerable level. In all three settings the additional uncertainty of the FH.MI estimator can be satisfactorily addressed and the bias is within an acceptable range. To have a closer look on the performance of the adapted Prasad-Rao MSE estimator the estimated and true RMSE values per domain are plotted in Figure 2.1 for the *mean* setting. First we observe that within each non-response rate the estimated RMSE decreases with higher sample size, which is in line with the behaviour of the true RMSE. Secondly, we see that per domain the estimated RMSE values increase with increasing non-response rates, which is consistent with the expected behaviour. At a non-response rate of 10% and 30%, the estimated RMSE tracks very well the behaviour of the true RMSE. With a higher non-response rate of 50% we see that there are underestimations in some areas, but overall the uncertainty is well accounted for. The proposed methods are good at capturing the additional variation due to the missing observations and imputation and also provide a realistic estimate of the uncertainty of the FH.MI estimator in our settings.

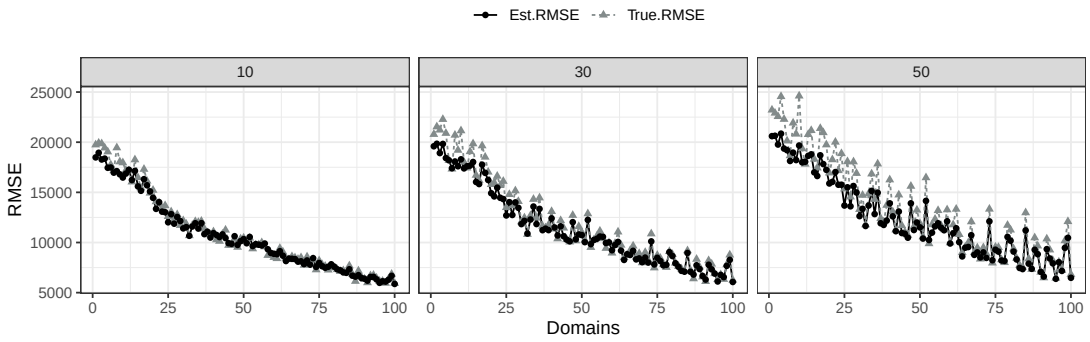


Figure 2.1: RMSE of FH.MI estimator per domain for *mean* setting and varying non-response rates. Domains are ordered by increasing sample size.

Table 2.3: Relative bias and RRMSE for mean and log mean, bias and RMSE for ratio

Non-response rate		before deletion		10%		30%		50%	
	Estimator	Mean	Median	Mean	Median	Mean	Median	Mean	Median
mean									
(rel.) Bias [%]	Direct	0.0464	0.0149						
	Direct.RR			0.0254	0.0198	0.0390	0.0092	0.0862	0.0290
	FH	0.2390	0.1812						
	FH.Direct.RR			0.2291	0.1691	0.2536	0.1872	0.3082	0.2583
	FH.MI			0.2245	0.1615	0.2355	0.1761	0.2704	0.2186
	FH.RR			0.2171	0.1568	0.2195	0.1639	0.2554	0.1840
log mean									
(rel.) Bias [%]	Direct	−0.2191	−0.0318						
	Direct.RR			0.1548	0.0342	1.1479	0.8566	2.8100	2.2903
	FH	−0.8797	−0.6057						
	FH.Direct.RR			0.0191	0.2091	1.4284	1.4639	3.1864	2.9609
	FH.MI			−0.2772	−0.1096	0.8383	0.8272	2.4169	2.3568
	FH.RR			−0.6948	−0.4258	0.0216	0.2115	1.3747	1.4485
ratio									
Bias	Direct	−0.0004	0.0000						
	Direct.RR			−0.0003	0.0000	−0.0000	0.0005	0.0009	0.0007
	FH	−0.0027	−0.0022						
	FH.MI			−0.0016	−0.0010	0.0012	0.0012	0.0011	0.0016
	FH.RR			−0.0026	−0.0021	−0.0024	−0.0018	−0.0015	−0.0009
mean									
RRMSE [%]	Direct	5.0318	4.2722						
	Direct.RR			5.1345	4.4849	5.5337	4.7889	6.1003	5.4419
	FH	4.4300	3.9609						
	FH.Direct.RR			4.5470	4.1570	4.9845	4.5694	5.6775	5.3471
	FH.MI			4.5444	4.1524	4.9643	4.5509	5.6018	5.2498
	FH.RR			4.5386	4.1385	4.9517	4.5388	5.5741	5.1978
log mean									
RRMSE [%]	Direct	25.5219	23.0001						
	Direct.RR			24.8037	22.0991	26.3014	23.1315	29.1076	26.1128
	FH	20.7739	20.0316						
	FH.Direct.RR			21.9160	21.4101	23.8789	22.5175	27.0243	25.9548
	FH.MI			21.3353	20.6552	22.7919	21.3174	25.4294	23.9328
	FH.RR			20.7741	19.9455	22.1078	20.6367	24.7177	23.3957
ratio									
RMSE	Direct	0.0655	0.0563						
	Direct.RR			0.0655	0.0565	0.0663	0.0565	0.0702	0.0617
	FH	0.0539	0.0506						
	FH.MI			0.0544	0.0510	0.0572	0.0533	0.0636	0.0607
	FH.RR			0.0541	0.0507	0.0564	0.0524	0.0624	0.0590

Table 2.4: Relative bias [%] of estimated RMSE (RBRMSE) of FH.MI

Non-response rate	10%		30%		50%	
	Mean	Median	Mean	Median	Mean	Median
mean	−1.4198	−1.8291	−3.4427	−3.1477	−6.9352	−6.9390
log mean	2.5119	2.5719	1.8185	2.6527	−4.0788	−3.2214
ratio	2.9396	3.1787	8.7815	9.0866	8.1231	8.2787

2.6 Application to Eurosystem's HFCS

In the following, we provide an example of how the proposed framework can be used for surveys with multiply imputed data in combination with small area methods. The purpose is to show a possible application with the HFCS data for scientists or institutions from relevant research areas rather than to discuss the estimates for each country. The HFCS is a large-scale survey of the financial and consumption situation of European households. The first wave was carried out in 2010 in 15 countries of the European Union (EU). The HFCS contains household data on both economic and demographic variables such as income, wealth, private pension, employment and consumption characteristics (Household Finance and Consumption Network, 2020a). So far three waves have been carried out, the last of which was collected in 2017 and released in March 2020. For the application the third wave is considered. The sample contains about 91,200 households in 22 countries of the EU, between 1,000 and 14,000 households per country. The HFCS is a joint project of several national statistical institutes, Eurosystem national central banks (NCB) and three non-euro area NCBs (Poland, Hungary, Croatia). For these countries, all values are converted into euros by the HFCN (Household Finance and Consumption Network, 2020a). The HFCN asked very sensitive questions, so the item non-response rate is high. Missing values in the HFCS data were iteratively and sequentially imputed. The variables are imputed along a path of imputation models. Each model is run several times, and the imputed values from the previous round are treated as given in the subsequent iteration (Household Finance and Consumption Network, 2020a). For each missing observation the HFCS data set contains $M = 5$ imputed values. For more information on the imputation method see Household Finance and Consumption Network (2020a). Of interest for this application is the value of the household's bonds, which is part of the household's assets and therefore relevant when considering the distribution of wealth. The HFCN reports conditional medians for the value of bonds per EU country (Household Finance and Consumption Network, 2020b). The values are calculated conditioned on households that have bonds; households with no bonds are discarded from the analysis. This results in partly very small sample sizes even on a country level, so that for some countries with fewer than 25 observations direct estimates are not reported by the HFCN. Furthermore, the rate of collected values differs between the countries. Since some households do not even indicate whether they own bonds or not, these values are also imputed by the HFCN. Therefore, the sample size per country, i.e., the number of households with bonds and the collected rate for these households, may differ slightly among the five imputed data sets provided by the HFCN. We calculate the sample sizes and collection rates based on the first imputed data sets. An overview of the sample sizes per country and the collected rates are given in Table 2.5. As dependent variable we choose the mean value of bonds in thousand of euros (TEUR) on a country level, resulting in $D = 22$ domains. In 2017 the EU consisted of 28 member states. 6 EU members are not included in the HFCS as their non-euro area NCBs do not participate. These domains are considered as out-of-sample (OOS) and model-based estimates are provided in the application. The direct estimators of the mean value of bonds for each imputed data set $\hat{\theta}^{Dir}_{d,m}$, $d = 1, ..., 22$, $m = 1, ..., 5$ are calculated according to Equation (2.13) using the sampling weights provided by the data provider, which corrects for potential bias due the sampling design and unit non-response. The

variances $\sigma^2_{e_{d,m}}$ are estimated with a bootstrap method following the instructions by Household Finance and Consumption Network (2020a) using the provided replicate weights derived by the Rao-Wu rescaled bootstrap method. As a result we obtain $M = 5$ replicates of direct estimators and their variances, which are then pooled according to Sections 2.3 and 2.4.

Table 2.5: Summary of EU-countries sample sizes, collected rates and auxiliary variables.

	Min	1stQ	Median	Mean	3rdQ	Max
Sample size	2.00	12.25	61.50	148.73	209.50	832.00
Collected rate	0.04	0.49	0.66	0.61	0.81	1.00
Total receipts from taxes and social contributions [% of GDP]	23.20	32.83	36.90	36.84	41.85	48.10
Final consumption expenditure [Current prices, EUR per capita]	5630	11710	17170	20424	29258	48140

2.6.1 Model selection and validation

To obtain auxiliary information from additional sources needed for the Fay-Herriot models, country-level data were collected from Eurostat, the statistical office of the EU and the European Commission. Within this set, data such as real estate data, unemployment rates, age dependency ratios, national accounts and tax aggregates from 2011 and 2017 were collected. The sources and years of this supplemental information are shown in Table B.1 in the Appendix. Due to the small number of domains, variables that were not available for the entire set of domains were excluded. The remaining auxiliary information includes variables such as the old, youth and age dependency ratio, the unemployment rate, the ratio of taxes to GDP, final consumption expenditure, the share of consumption expenditure on GDP, GDP at market prices and a variable indicating whether the country has a wealth tax. In addition, the number of covariates in the model is severely limited by the small number of domains, which is why we restricted the model to two possible auxiliary variables. In the context of area-level data, Han (2013) transferred the conditional Akaike information in linear mixed models from Vaida and Blanchard (2005) to a conditional Akaike information criterion for Fay-Herriot models. Marhuenda et al. (2014) examine this criterion among Kullback symmetric divergence criterion (KIC) and propose a bootstrap variant of the KIC (KICb2) especially developed for FH models. They conclude that KICb2 criterion is one of the best model selection criteria for Fay-Herriot models. Therefore, in this application the preselection of variables was performed using the KICb2 criterion. Model selection was carried out for each of the 5 imputed datasets, with no particular difference in the results. A union of two auxiliary variables was selected for the final model, as shown in Table 2.5. To obtain a model-based estimator of the mean value of household bonds, the estimator from Section 2.4.1 is calculated with the auxiliary information in Table 2.5. The model variances σ_v^{2RR} are calculated for the MI-adjusted Fay-Herriot model on the original scale using the REML method. The distributional assumptions of the model presented in Section 2.3 are checked by the Shapiro-Wilk test applied to the residuals and the random effects. For the MI-adjusted Fay-Herriot model for a normal mean, the p-values of the tests for the standardized residuals and the random effect are 0.223 and 0.965, respectively. Therefore, the normality assumptions for both error terms cannot be rejected at a 5% signifi-

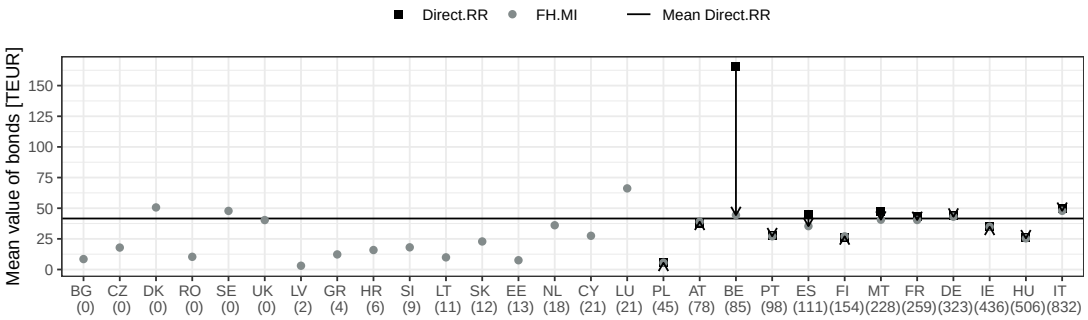


Figure 2.2: Direct and model-based estimates for the mean value of bonds, own estimations. Domains are ordered by increasing sample size, sample sizes in brackets. Direct estimates for domains with less than 25 observations are not reported.

cance level. Consequently, all further considerations and results are based on the MI-adjusted Fay-Herriot model for a mean value as presented in Section 2.4.1. The explanatory power of the model is assessed using the modified R^2 for Fay-Herriot models according to Lahiri and Suntornchost (2015) and we obtain a value of 45%. Due to the low number of domains, it is not possible to include more auxiliary variables to potentially increase explanatory power. We obtain positive estimated regression coefficients for both auxiliary variables. The impact on the tax-to-GDP ratio seems reasonable, given that tax contributions include taxes on wealth (at least in some countries) and that high tax revenues from income could indicate a high level of capital assets. The relationship between consumption and wealth is not independent of income, because if income is higher than consumption, the rest can be invested, and if consumption cannot be covered by income, there is nothing left to invest. Nevertheless, with the given data, the model also shows a positive effect for consumption.

2.6.2 Small area estimates

The estimates of the mean value of bonds on a country level are calculated using the FH.MI estimator for a mean value and to estimate the MSE the MI adapted Prasad-Rao estimator is applied as described in Section 2.4.1. To compare the model-based estimators with a direct estimator, the direct estimators and their variance estimates are computed for each imputed data set as described above and pooled using Rubin's rule in Equation (2.8) (Direct.RR). The point estimates of the model-based estimators (FH.MI) should be consistent with the unbiased estimates of the direct estimator, but be more precise. Figure 2.2 compares the direct and the model-based point estimates for the 22 in-sample domains and additionally reports the estimates for the 6 OOS EU countries. Due to the guidelines of the data provider, the direct estimates for domains with less than 25 observations are not reported. We observe that, for countries with large sample sizes, the direct and model-based estimates are almost identical, consistent with the expectation that high weight is given to the direct estimator when precision is high. An exception is Belgium (BE), where the sample size is rather high, but the shrinkage to the mean quite strong. For most of the direct estimates, which tend to be high, we see that the model-based estimates are smaller, showing the shrinkage effect to the mean of the

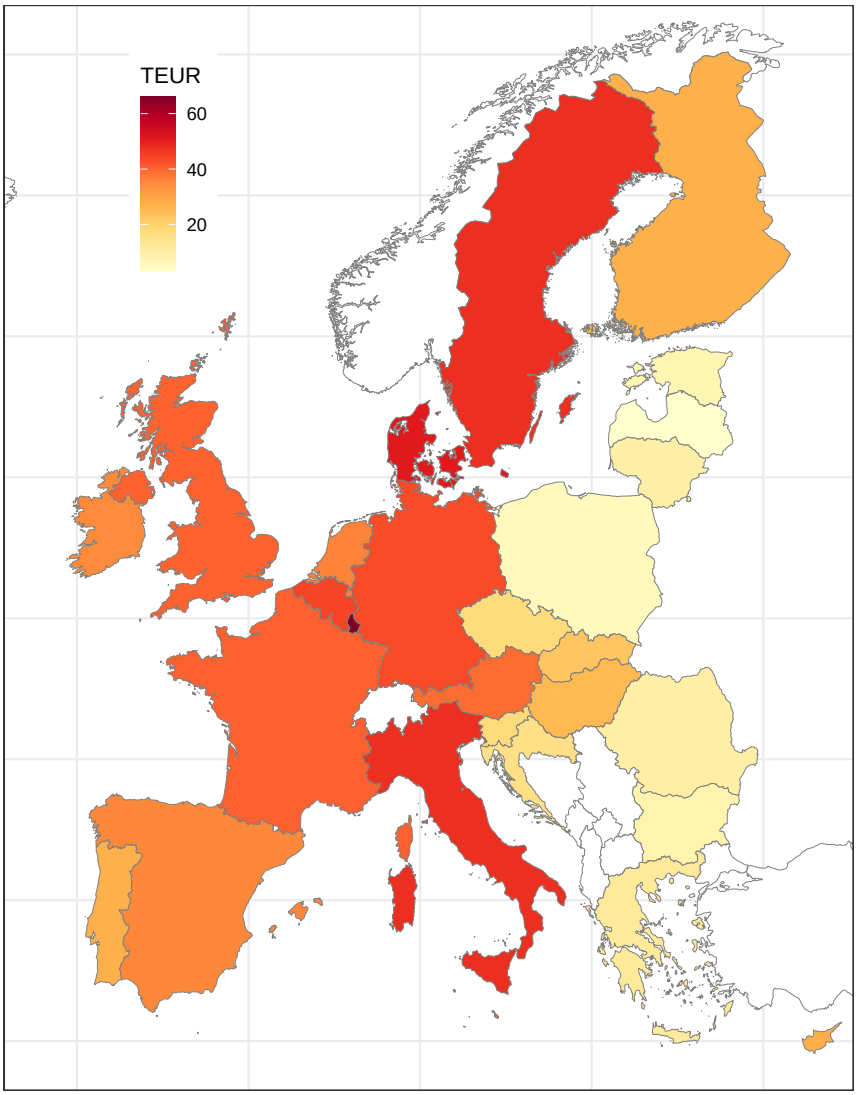


Figure 2.3: Map of model-based FH.MI estimates for mean value of bonds, own estimations. Non-EU countries in 2017 are colored in white.

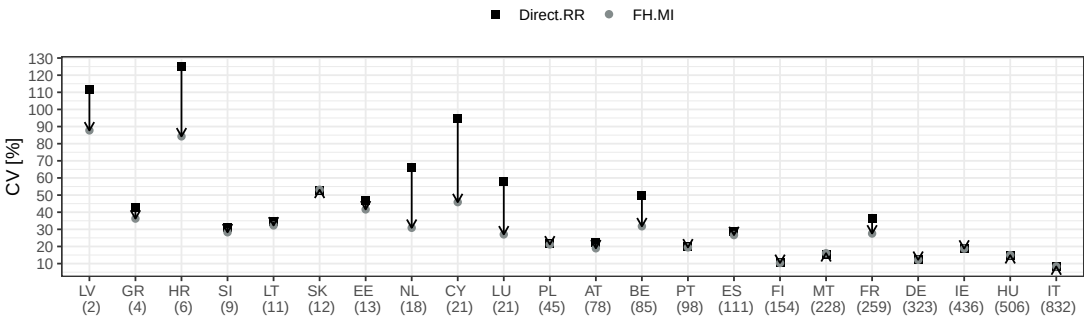


Figure 2.4: CVs of direct and model-based estimates, own estimations. Domains are ordered by increasing sample size, sample sizes in brackets.

model-based estimates (see summary statistics of point estimates in Table B.2 in the appendix). Possibly due to the low number of covariates very little shrinkage takes place for some countries with small sample sizes (GR, SI, LI). The model-based point estimators are furthermore reported in the map in Figure 2.3. The highest values are estimated for Luxembourg (LU), followed by Denmark (DK) (OOS) and Sweden (SE) (OOS). For eastern European countries, the estimates are rather low, followed by southern European countries. The estimated model-based values range from 3 to 66 thousand euros (cf. Table B.2 in the appendix), which seems plausible given the median values reported by the HFCN (Household Finance and Consumption Network, 2020b) between 2 and 25 thousand euros, considering that the distribution at the household level tends to be right skewed and therefore the mean values should be higher than the median values. Figure 2.4 shows the coefficients of variation (CV) for the direct and model-based estimates. We see that the model-based estimator is at least as efficient as the direct estimator. The CVs of the model-based estimators are mostly significantly smaller than those of the direct estimators, with the effect decreasing with increasing sample size. For large sample sizes, the gain is barely noticeable, but this is consistent with the expected behavior that the direct estimator is sufficiently accurate in this case. For some domains, such as Croatia (HR) and Cyprus (CY), the CV is almost halved. Due to the relatively small domain size of $D = 22$ and hence the limitation to the number of covariates in the model, the efficiency gain is limited. A summary of the distribution of the point estimators and CVs from Figures 2.2 and 2.4 can be found in Table B.2 in the appendix.

2.7 Concluding remarks

In this paper, we derive small area indicators based on multiply imputed survey data and present uncertainty measures for common cases that capture the additional uncertainty. We present the transformed Fay-Herriot model calculated on each imputed data set. We then combine the components into a MI adjusted Fay-Herriot model that retains the model structure of the Fay-Herriot model. With this approach, results that exist for the Fay-Herriot model regarding transformations, back-transformations and MSE estimators can be extended. It is a general approach that can be applied to any indicator with a given transformation and an appropriate back-transformation. We discuss common special cases of the model (mean, log mean, arcsine

ratio). For these special cases we propose MSE estimators. For the mean and logarithmic mean, we present an analytical adaption of the Prasad-Rao estimator and, for the arcsine ratio, we use a bootstrap estimator. We demonstrate in simulation studies that the resulting FH.MI point estimators lead to valid results in terms of bias and RMSE in the given settings and under different non-response rates and that the proposed MSE estimators are able to capture the additional imputation uncertainty and lead to good uncertainty measures. We carried out an application using the proposed framework to obtain estimates for European household assets.

A limitation of the proposed approach is that it is not as straightforward for the user as it would be if only the Fay-Herriot estimators were estimated for each imputed data set and the mean value calculated. But, as mentioned above, it is not clear how the variance pooling rules can be applied to the MSE. This could be part of further research. To facilitate the application, it is planned to provide an `R`-package with the methodology presented. Other open research questions are the extension from a cross-sectional to a longitudinal analysis to provide stable estimates across panel waves (i.e., over time) when multiple imputations are performed and sample sizes are small. If the underlying data structure is a panel survey and individuals or households are observed over multiple time periods, the Fay-Herriot model can be adapted to consider the correlation of the same observations over time. To borrow strength for domain estimates, Rao and Yu (1994a) propose a model with auto-correlated random effects and assume an auto-regressive process of first order. In addition to the temporal Fay-Herriot models, a multivariate approach could serve the requirement to consider the temporal dimension in the data. In the multivariate Fay-Herriot model (Benavent and Morales, 2016) the domain indicators are estimated simultaneously for the different panel waves. In this way, correlations for both error terms can be considered. These models have not yet been investigated in combination with multiple imputation. The approach in this paper could be extended to include correlations over time to ensure reliable estimates over time based on multiply imputed survey data. Since asset values are usually highly skewed, more robust indicators such as the median or other quantiles could be estimated instead of the mean. Therefore, the estimation of small area medians using the Fay-Herriot model would be interesting for future research.

Acknowledgements

The authors appreciate gratefully the support of the German Research Foundation within the TESAP project (grant number: 281573942). This article uses data from the Eurosystem Household Finance and Consumption Survey (HFCS). The results published and the related observations and analysis may not correspond to results or analysis of the data producers. Finally, the authors are indebted to the Editor-in-Chief, Associate Editor and the referees for comments that significantly improved the article.

Appendix B

A.1 MSE back-transformation for a log mean

Let $\mu = exp(\theta)$ be the true parameter value and $\hat{\theta}$ be an estimate for θ. Furthermore, $\hat{\mu}$ is an estimator for μ with $\hat{\mu} = g(\hat{\theta})$, where g is a continuously differentiable function. For

$$g(\hat{\theta}) = exp\{\hat{\theta} + 0.5\widehat{\text{MSE}}(\hat{\theta})\}$$

an approximation of $\text{MSE}(\hat{\mu})$ using Taylor expansion can be derived as follows:

$$\begin{aligned}
\text{MSE}(g(\hat{\theta})) &= \text{Var}(g(\hat{\theta})) + \text{Bias}^2(g(\hat{\theta})) \\
&= \text{E}[g(\hat{\theta})^2] - \text{E}[g(\hat{\theta})]^2 + \text{E}[g(\hat{\theta}) - g(\theta)]^2 \\
&\approx \text{E}[\{g(\theta) + g'(\theta)(\hat{\theta} - \theta)\}^2] - \text{E}[\{g(\theta) + g'(\theta)(\hat{\theta} - \theta)\}]^2 + \text{E}[g'(\theta)(\hat{\theta} - \theta)]^2 \\
&= g'(\theta)^2\{\text{E}[\hat{\theta}^2] - \text{E}[\hat{\theta}]^2\} + g'(\theta)^2\text{E}[\hat{\theta} - \theta]^2 \\
&= g'(\theta)^2\{\text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta})\} = g'(\theta)^2\text{MSE}(\hat{\theta}).
\end{aligned}$$

A estimator of $\text{MSE}(\hat{\mu})$ is then obtained by

$$\widehat{\text{MSE}}(\hat{\mu}) = \widehat{\text{MSE}}(g(\hat{\theta})) = g'(\hat{\theta})^2\widehat{\text{MSE}}(\hat{\theta}) = exp\left\{\hat{\theta} + 0.5\widehat{\text{MSE}}(\hat{\theta})\right\}^2 \widehat{\text{MSE}}(\hat{\theta}).$$

A.2 Plots and tables

	Year	Source
Private households by type, tenure status (Real estate)	2011	Eurostat (2011b)
Dwellings by occupancy status, type of building (Real estate)	2011	Eurostat (2011a)
Age, Old, Young-age dependency ratios	2017	Eurostat (2017d)
Unemployment rate	2017	Eurostat (2017a)
Tax to GDP ratio	2017	Eurostat (2017c)
Final consumption expenditure	2017	Eurostat (2017b)
GDP at market prices	2017	Eurostat (2017b)
Share of consumption expenditure on GDP	2017	Eurostat (2017b)
Indicator for presence of wealth tax	2017	European Commission (2017)

Table B.1: Source and year of auxiliary information.

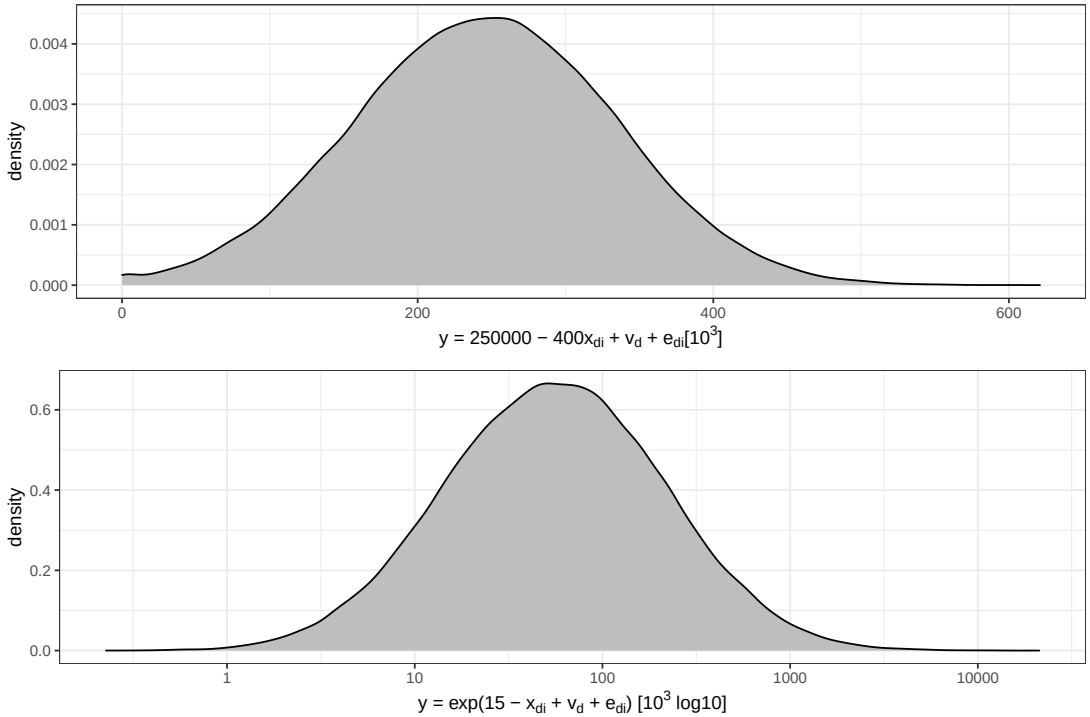


Figure B.1: Density of population target variable of one replication

Table B.2: Summary of point estimators and CVs for mean value of bonds [TEUR].

Estimator		Min	1stQ	Median	Mean	3rdQ	Max
Direct.RR	Point est.	2.5	19.6	36.2	41.6	49.0	165.5
FH.MI		3.1	16.4	27.3	28.6	40.0	66.2
Direct.RR	CV [%]	8.3	19.3	32.8	41.9	51.7	125.0
FH.MI		8.3	18.8	27.3	31.5	35.3	87.8

Part II

Optimal Transformations and Model Building for Estimating Regional Indicators

Chapter 3

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

This is the peer reviewed version of the following article: Lee, Y., Rojas-Perilla, N., Runge, M. and Schmid, T. (2023) Variable selection using conditional AIC for linear mixed models with data-driven transformations. *Statistics and Computing 33(27)*, which has been published in final form at https://doi.org/10.1007/s11222-022-10198-9.

3.1 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, 2010) 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 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 a 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 3.2 we provide an overview of linear mixed models and the $cAIC$. In Section 3.3, we derive the Jacobian adjusted $cAIC$ for transformed linear mixed models and introduce the step-wise selection approach. In Section 3.4, we examine the performance of the proposed selection approach by using model-based simulations. In Section 3.5, the proposed selection approach is applied to data from Guerrero in Mexico for estimating poverty and inequality indicators at the municipal level. Finally, we discuss our results and further directions of research in Section 3.6.

3.2 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 3.2.1, we present a general notation of linear mixed models and in Section 3.2.2, we introduce and compare the $cAIC$ by Vaida and Blanchard (2005) and Liang et al. (2008).

3.2.1 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

$$y = X\beta + Zu + \varepsilon, \tag{3.1}$$

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

$$Cov(y) = V = \sigma^2 I_N + Z G_0 Z^T.$$

3.2.2 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 (3.1) 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

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

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, 2010). 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)

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

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

$$\begin{aligned} K_{MLE} =& \frac{N(N-p-1)}{(N-p)(N-p-2)}(\rho + 1) \\ &+ \frac{N(p+1)}{(N-p)(N-p-2)}. \end{aligned} \tag{3.2}$$

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.

3.3 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 3.3.2, 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 3.3.3., 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 3.3.4.

3.3.1 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

$$T_\lambda(y_{ij}) = \begin{cases} \frac{(y_{ij}+s)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(y_{ij}+s) & \text{if } \lambda = 0, \end{cases} \tag{3.3}$$
$$i = 1, ..., D \text{ and } j = 1, ..., N_i,$$

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

$$T_\lambda(y) = \widetilde{y} = X\beta + Zu + \varepsilon \tag{3.4}$$

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

$$T_{\lambda_m}(y) = \widetilde{y}^{(m)} = X^{(m)}\beta + Zu + \varepsilon, \tag{3.5}$$
$$m = 1, ..., M,$$

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 (3.5) 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, 2010). Therefore, an adjustment with the Jacobian to the $cAIC$ should be performed first such that these M different models can be compared.

3.3.2 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

$$\begin{aligned} BC =& - E_{f(\widetilde{y},u)}[-\log g(\widetilde{y}|\hat{\theta}, \hat{u})] + 0.5 \cdot cAI \\ =& E_{f(\widetilde{y},u)}[\log g(\widetilde{y}|\hat{\theta}, \hat{u})] \\ & - E_{f(\widetilde{y},u)} E_{f(\widetilde{y}^*|u)}[\log g(\widetilde{y}^*|\hat{\theta}, \hat{u})] \\ =& E\left[\frac{1}{2\sigma^2}[(\widetilde{y}^* - \widehat{\widetilde{y}})^T(\widetilde{y}^* - \widehat{\widetilde{y}}) - (\widetilde{y} - \widehat{\widetilde{y}})^T(\widetilde{y} - \widehat{\widetilde{y}})]\right], \end{aligned} \tag{3.6}$$

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

Under the assumption that σ^2 is unknown, the BC in Equation (3.6) can be replaced by K_{MLE} from Equation (3.2). Consequently, the $cAIC$ for the transformed model is given by

$$cAIC = -2 \log g(\widetilde{y}|\hat{\theta}, \hat{u}) + 2K_{MLE}, \tag{3.7}$$

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 (3.7) 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

$$\begin{aligned} h(y|u_0) =& f(\widetilde{y}|u_0) \cdot J(\lambda, y), \\ l(y|\theta, u) =& g(\widetilde{y}|\theta, u) \cdot J(\lambda, y), \end{aligned} \tag{3.8}$$

where

$$J(\lambda, y) = \left|\frac{\partial \widetilde{y}}{\partial y}\right| = \prod_{i=1}^{D}\prod_{j=1}^{N_i} \frac{\partial \widetilde{y}_{ij}}{\partial y_{ij}} = \prod_{i=1}^{D}\prod_{j=1}^{N_i} (y_{ij} + s)^{\lambda - 1}. \tag{3.9}$$

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

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

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):

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

$l(y|\hat{\theta}, \hat{u})$ is defined as in Equation (3.8). 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 (3.8). By inserting these terms into the BC, we get

$$\begin{aligned} BC =& E_{h(y,u)}[\log(g(\widetilde{y}|\hat{\theta}, \hat{u}) \cdot J(\hat{\lambda}, y))] \\ & - E_{h(y,u)} E_{h(y^*|u)}[\log(g(\widetilde{y}^*|\hat{\theta}, \hat{u}) \cdot J(\hat{\lambda}, y^*))] \end{aligned}$$

$$
\begin{aligned}
&= E\Big[-\frac{N}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\widetilde{y}-\widehat{\widetilde{y}})^T(\widetilde{y}-\widehat{\widetilde{y}}) \\
&\qquad + \log(J(\hat{\lambda}, y))\Big] \\
&\quad -E\Big[-\frac{N}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\widetilde{y}^*-\widehat{\widetilde{y}})^T(\widetilde{y}^*-\widehat{\widetilde{y}}) \\
&\qquad + \log(J(\hat{\lambda}, y^*))\Big]. \qquad (3.10)
\end{aligned}
$$

The Jacobian term of y is defined in Equation (3.9) 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

$$
\begin{aligned}
BC = E\Big[&-\frac{N}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\widetilde{y}-\widehat{\widetilde{y}})^T(\widetilde{y}-\widehat{\widetilde{y}}) \\
&+ (\hat{\lambda}-1)\sum_{i=1}^{D}\sum_{j=1}^{N_i}\log(y_{ij}+s))\Big] \\
-E\Big[&-\frac{N}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\widetilde{y}^*-\widehat{\widetilde{y}})^T(\widetilde{y}^*-\widehat{\widetilde{y}}) \\
&+ (\hat{\lambda}-1)\sum_{i=1}^{D}\sum_{j=1}^{N_i}\log(y_{ij}^*+s)\Big]. \qquad (3.11)
\end{aligned}
$$

3.3.3 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 (3.11) 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 (3.4) 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

$$
\begin{aligned}
BC = \frac{1}{B}\sum_{b=1}^{B}\Bigg[& -\frac{N}{2}log\big(2\pi\hat{\sigma}^{2(\hat{\lambda}^{(b)},(b))}\big) - \frac{1}{2\hat{\sigma}^{2(\hat{\lambda}^{(b)},(b))}}\cdot \\
& \left(\widetilde{y}^{(\hat{\lambda}^{(b)},(b))} - X\hat{\beta}^{(\hat{\lambda}^{(b)},(b))} - Z\hat{u}^{(\hat{\lambda}^{(b)},(b))}\right)^{T} \\
& \left(\widetilde{y}^{(\hat{\lambda}^{(b)},(b))} - X\hat{\beta}^{(\hat{\lambda}^{(b)},(b))} - Z\hat{u}^{(\hat{\lambda}^{(b)},(b))}\right) \\
& + \big(\hat{\lambda}^{(b)} - 1\big)\sum_{i=1}^{D}\sum_{j=1}^{N_i}\log\big(y_{ij}^{(b)} + s\big)\Bigg] \\
-\frac{1}{B}\sum_{b=1}^{B}\Bigg[& -\frac{N}{2}log\big(2\pi\hat{\sigma}^{2(b)}\big) - \frac{1}{2\hat{\sigma}^{2(b)}}\cdot \\
& \left(\widetilde{y} - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\right)^{T} \\
& \left(\widetilde{y} - X\hat{\beta}^{(b)} - Z\hat{u}^{(b)}\right) \\
& + \big(\hat{\lambda} - 1\big)\sum_{i=1}^{D}\sum_{j=1}^{N_i}\log\big(y_{ij} + s\big)\Bigg]. \qquad (3.12)
\end{aligned}
$$

Then, the $JcAIC$ is estimated by

$$
\begin{aligned}
JcAIC = & -2\log(l(y|\hat{\theta},\hat{u})) + 2BC \\
= & -2\log(g(\widetilde{y}|\hat{\theta},\hat{u})) - 2\log(J(\hat{\lambda},y)) \\
& + 2BC \qquad (3.13)
\end{aligned}
$$

with BC defined in Equation (3.12).

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.

3.3.4 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 3.4 and the application in Section 3.5.

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 (3.13) 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.

3.4 Model-based simulation experiment

To support our theoretical findings and the proposed framework from the previous section, we conduct simulation studies that include several settings. The aim of the 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 3.1. 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 $N = 10,000$ 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 C.1 and Figure C.1.

In addition to the explanatory variables $x_{1,ij}$, $x_{2,ij}$ 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):

$$\begin{aligned} &T_\lambda(y_{ij}) = \widetilde{y}_{ij} = \\ &\beta_0 + \beta_1 x_{1,ij} + \beta_2 x_{2,ij} + \beta_3 x_{3,ij} + \beta_4 z_{ij} + u_i + e_{ij}, \end{aligned} \tag{3.14}$$

where T denotes the Box-Cox transformation defined in Equation (3.3). In each simulation run, the model selection is performed with four approaches, where the dependent variable y is on different scales:

Table 3.1: 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)	$400 - 10x_{1,ij} + 100x_{2,ij} - 10x_{3,ij} + u_i + e_{ij}$	$\mathcal{N}(\mu_i, 3^2)$	$\mathcal{U}[-3, 3]$	$\text{Bin}(1, 0.8)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 30^2)$	$\mathcal{N}(0, 60^2)$
Normal (2)	$400 - 10x_{1,ij} + 100x_{2,ij} - 10x_{3,ij} + u_i + e_{ij}$	$\mathcal{N}(\mu_i, 3^2)$	$\mathcal{U}[-3, 3]$	$\text{Bin}(1, 0.8)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 10^2)$	$\mathcal{N}(0, 20^2)$
Log	$exp(10 - x_{1,ij} + x_{2,ij} - 0.5x_{3,ij} + u_i + e_{ij})$	$\mathcal{N}(\mu_i, 2^2)$	$\mathcal{U}[2, 3]$	$\text{Bin}(1, 0.8)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 0.4^2)$	$\mathcal{N}(0, 0.8^2)$
Box-Cox	$[(10 - x_{1,ij} + x_{2,ij} - 0.5x_{3,ij} + u_i + e_{ij})(-0.5) + 1]^{\frac{1}{-0.5}}$	$\mathcal{N}(\mu_i, 2^2)$	$\mathcal{U}[2, 3]$	$\text{Bin}(1, 0.8)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 0.4^2)$	$\mathcal{N}(0, 0.8^2)$

- 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 3.3.4. For each setting, the linear mixed model from (3.14) 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 (3.13) 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. 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, 2022). 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 3.2 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

Table 3.2: 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

Table 3.3: 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

$JcAIC$ values for *Box-Cox Opt* and *Box-Cox Naive* are identical. This makes sense considering the corresponding estimated λs in Table 3.3, 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 3.3) are close to zero, which results in a log transformation 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 3.4 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 3.3. 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 3.4, 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.

Table 3.4: 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

3.5 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 3.5.1, we provide a brief overview of the small area estimation and the EBP. In Section 3.5.2, 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 3.5.3.

3.5.1 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, 1994b; 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

$$y_{ij} = x_{ij}^T \beta + u_i + \varepsilon_{ij}, \tag{3.15}$$

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

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).

3.5.2 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 Estadistica y Geografia (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 3.5 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 (3.15) to explain `ictpc`, the Gaussian assumptions of random effects and errors are required. However, the histogram of `ictpc` in Figure 3.1 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 3.5: 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

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 3.3. 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 3.3.4 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

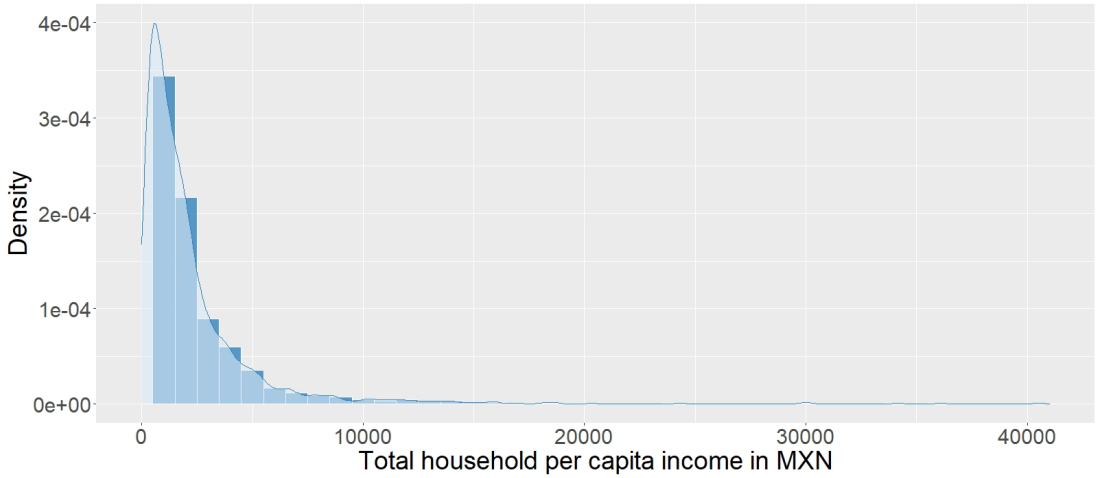


Figure 3.1: Distribution of the total household per capita income in MXN (`ictpc`)

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 3.3, *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.

3.5.3 Results

First, the chosen variables for the optimal model and the optimal transformation parameters of each approach are compared. Table 3.6 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.

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 3.7. 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.

Since the linear mixed model relies on Gaussian assumptions and we decided to use a transformation

Table 3.6: 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

Table 3.7: 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

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 3.8). 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 C.2 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 3.8: 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 3.2. Figure 3.2 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.

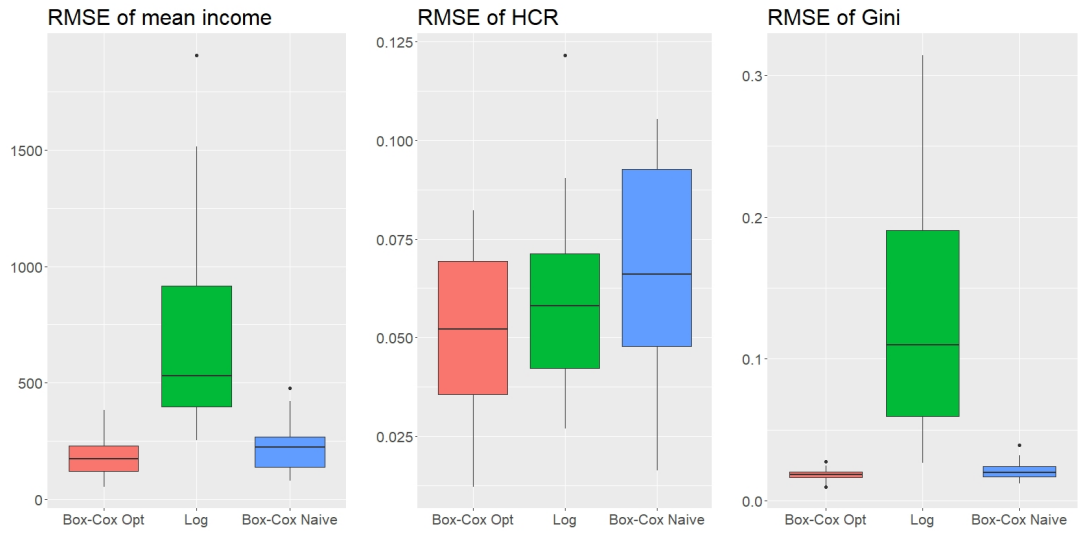


Figure 3.2: RMSE of EBP estimates for mean income, HCR and Gini

Figure 3.3 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.

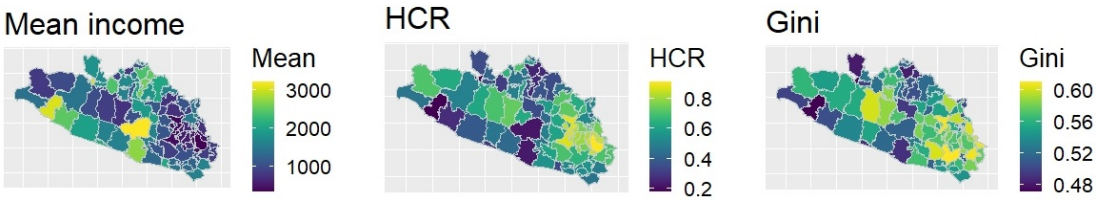


Figure 3.3: EBP estimates for mean income, HCR and Gini based on *Box-Cox Opt* approach.

3.6 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.

Appendix C

C.0.1 Bootstrap for *Original* and *Log* approach

A. Bootstrap for *Original*

1. Fit the model in Equation (3.1) 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. Refit 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

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

B. 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

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

with

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

C.0.2 Graphics and Tables

Table C.1: Summary statistics of the dependent variable (y_{ij}) in the first Monte Carlo population.

Data setting	Min	1Q	Median	Mean	3Q	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

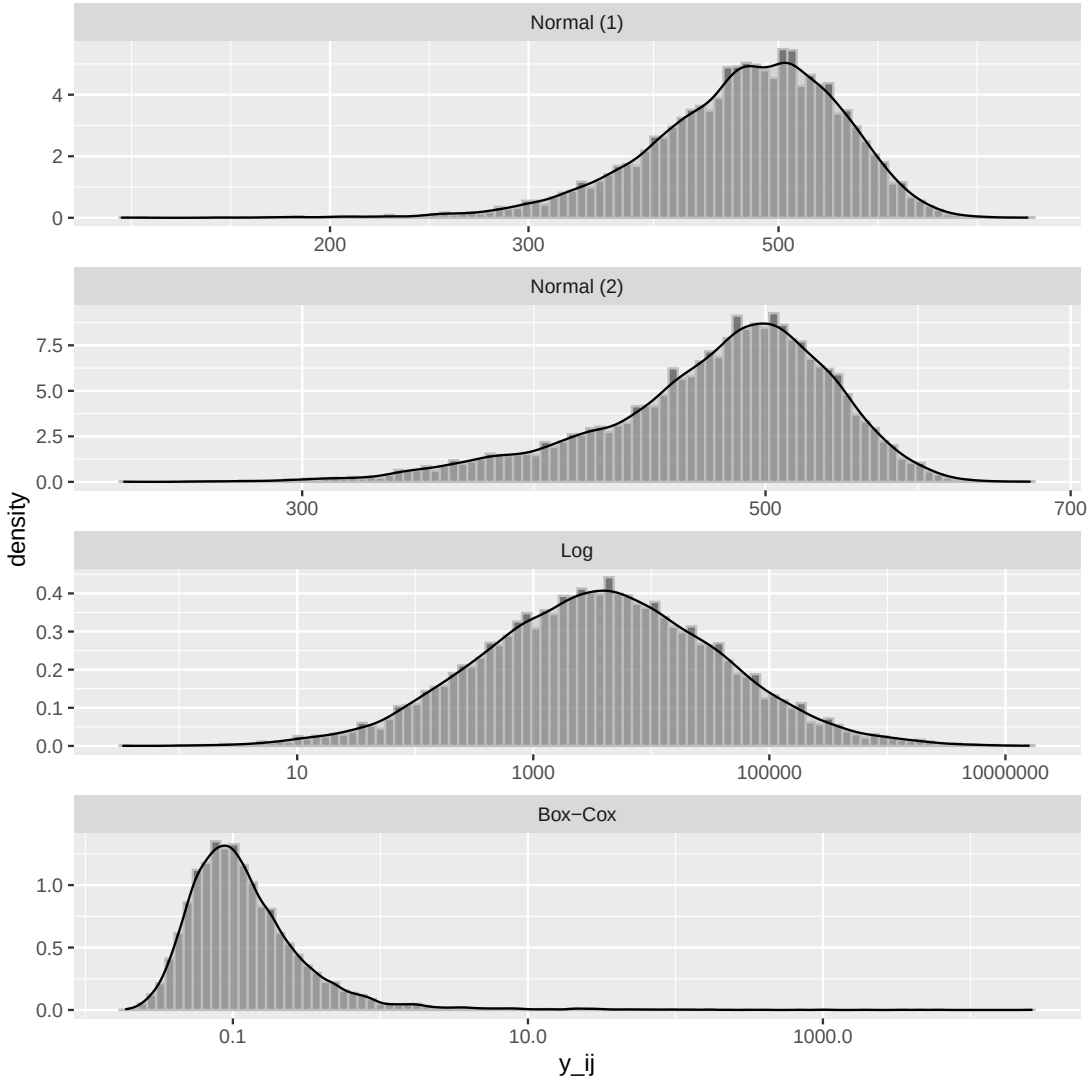


Figure C.1: Density of the dependent variable (y_{ij}) in the first Monte Carlo population. Note that a base-10 log scale is used for the x-axis for the *Log* and *Box-Cox* setting.

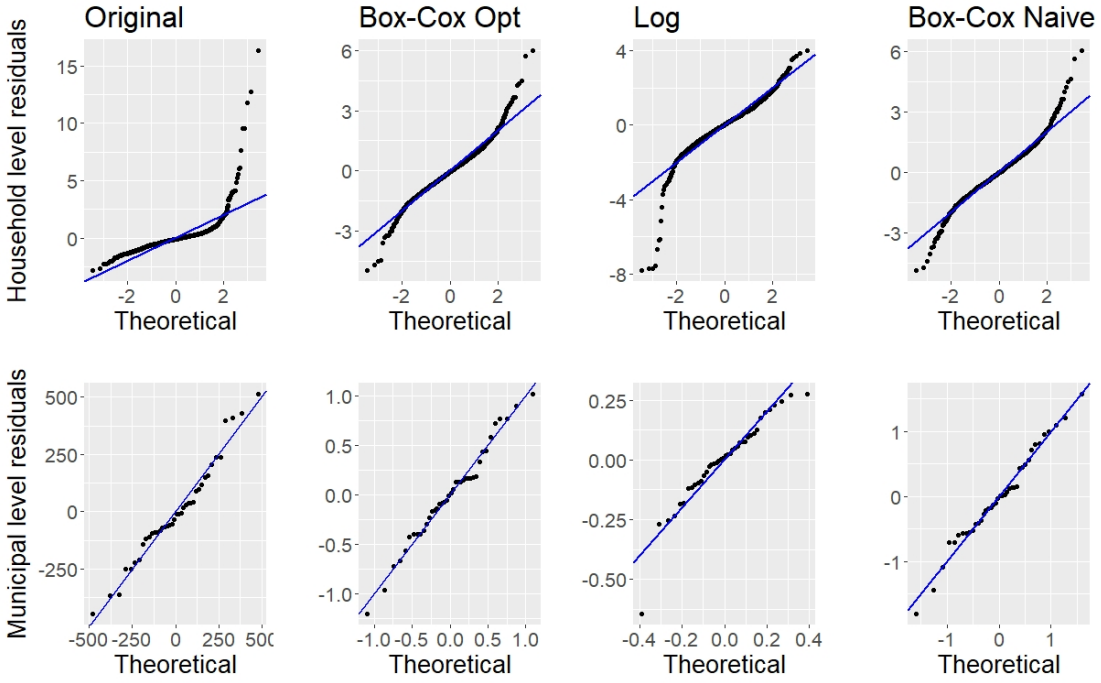


Figure C.2: Q-Q plots for household level and municipal level residuals of different EBP approaches.

Chapter 4

Area-level small area estimation with random forests

4.1 Introduction

For developing countries, the availability of socio-demographic indicators at a small regional level is particularly important for targeting policy interventions where they are most needed. Moreover, there are often large regional disparities within these countries. Increasing national growth rates and reducing regional disparities are crucial to achieving the United Nations Sustainable Development Goals, which aim to ensure sustainable economic, social and environmental development worldwide (United Nations, 2012). Spatially disaggregated poverty estimates therefore help to assess where the provision of public and social services is most important. With the help of small area estimation (SAE) methods reliable estimates of disaggregated indicators can be produced. Estimators for such disaggregated indicators that are derived only from the domain-specific survey data (hereafter referred to as direct estimators) tend to be unreliable because sample sizes may be insufficient and existing surveys are often designed for higher levels, such as the national level. Since the enlargement of the sample size of surveys is cost intensive and time consuming, SAE methods are developed to obtain estimates in a small area or domain with adequate precision. This is done by using model-based approaches and enriching survey data by additional information from further data sources traditionally taken from administrative data (census or register data). A small area or domain refers to any subpopulation of the population of interest, e.g. geographic areas or socio-demographic groups. SAE methods can be distinguished into two types: unit- and area-level models. While unit-level models relate the unit values of a variable of interest to auxiliary information at the individual level (Battese et al., 1988; Molina and Rao, 2010), area-level models use survey data and auxiliary information, both aggregated to the desired area-level (Fay and Herriot, 1979). For more detailed overviews of SAE methods, see Pfeffermann (2013); Rao and Molina (2015); Tzavidis et al. (2018) and Jiang and Rao (2020). Even though unit-level models may lead to a gain in precision since more information is used for the model estimation, area-level models offer a valuable alternative. Access to auxiliary information like census or register data is due to confidentiality reasons often less likely, but aggregated auxiliary information are provided more frequently. Additionally, area-level models can take the survey design into account by integrating the sampling weights into the direct estimation. Both types of models are based on linear mixed models (LMM), which offer a great opportunity to model area effects that are not explained by area-specific covariates due to the random effects of the model. To avoid biased estimates and unreliable mean squared error (MSE) estimates, the model assumptions of LMMs have to be met. The violation of model assumptions and other problems that might occur in practical SAE applications and possible solutions are exemplarily: To encounter

the violation of normally distributed random effects and error terms, transformations can be applied, e.g. a log transformation or parametric transformations like the Box-Cox or dual power transformation for right skewed data (Rao and Molina, 2015; Slud and Maiti, 2006; Sugasawa and Kubokawa, 2017). Robust area-level models are developed to handle influential outlying observations (Chambers et al., 2014; Schmid et al., 2016; Jiang and Rao, 2020). For the estimation of ratios the arcsine transformation guarantees that the model results lie in the interval $[0; 1]$ (Casas-Cordero et al., 2016; Schmid et al., 2017; Sugasawa and Kubokawa, 2017; Hadam et al., 2020). For applications where the functional form of the relationship between the response variable and the auxiliary variables is nonlinear or unknown, Giusti et al. (2012) develop a semiparametric Fay-Herriot (FH) model based on penalized splines. Another problem may be that unknown interactions between explanatory variables may affect the model estimation and in traditional (mixed) regression models interaction terms have to be included by the practitioner. One issue that arises in traditional SAE data applications is model building. Practitioners have to decide which auxiliary variables should be included in the model. This decision is often based on theoretical considerations and/or variable selection criteria like the Akaike or Bayesian information criterion or the R^2 of the resulting model. For area-level models, Marhuenda et al. (2014) derive bootstrap and bias corrected versions of the mentioned variable selection criteria and Lahiri and Suntornchost (2015) propose an adjusted R^2.

However, all the listed approaches only tackle single problems. Machine learning methods represent a meaningful alternative to encounter several problems at the same time. In this paper, the focus lies in particular on random forests (RF) (Breiman, 2001). The use of RFs combines many advantages: RFs are not limited to (parametric) model assumptions, they learn predictive relations from data, meaning that they are able to capture nonlinear relations between the response variable and the auxiliary variables and to handle higher order interactions between auxiliary variables (Hastie et al., 2008; Varian, 2014). RFs are known for an excellent predictive performance even when working with skewed data or facing influential outliers. Since RFs cannot extrapolate, the predictions of RFs automatically lie in a predefined range of values (dependent on the input data). In addition, RFs perform implicit variable selection (Biau and Scornet, 2016).

The goal of this paper is to combine the advantages of area-level SAE models and RFs. In the context of mixed models, there is the mixed effects random forest model (MERF) approach (Hajjem et al., 2014). The MERF combines the method of RFs while modeling the hierarchical dependencies of mixed models. In the field of SAE, Krennmair and Schmid (2022) adapt MERFs to unit-level SAE models. Conceptually, we transfer the methods introduced by Krennmair and Schmid (2022) to area-level models. Even for unit-level SAE models, few studies consider tree-based methods (e.g.Mendez (2008); Bilton et al. (2017); De Moliner and Goga (2018); McConville and Toth (2019)), and to the best of our knowledge, there is no literature yet for area-level models. We aim to fill this gap by introducing the area-level MERF and a corresponding nonparametric bootstrap MSE estimator to measure its uncertainty. The performance of the newly introduced method is investigated and compared to the standard and semiparametric FH estimators by model-based simulations. We illustrate the area-level MERF in an application based on remote-sensing data. Traditionally, FH models combine survey and census data, but there are several alternatives when no (recent) census data is available. For example, survey data (Ybarra and Lohr, 2008), general big data sources (Marchetti et al., 2015), Google Trends data (Porter et al., 2014), or mobile phone data (Schmid et al., 2017) have already been used in area-level SAE applications. Other valuable auxiliary information is geospatial data that stem from remote-sensing sources like satellite imagery. In the field of poverty estimation and estimation of socio-economic indicators, Seitz (2019) applies a FH model and uses auxiliary information derived from satellite imagery to estimate different welfare indicators for Central Asian districts. Newhouse et al. (2022) and Masaki et al. (2022) compare unit- and area-level models based on survey and satellite data in order to estimate monetary poverty for Mexican municipalities and non-monetary poverty in Sri Lanka and Tanzania,

respectively. Edochie et al. (2023) adapt geospatial auxiliary data for estimating poverty rates at the unit-level in Chad, Guinea, Mali and Niger, and additionally compare results at the unit- and area-levels for the country of Burkina Faso. To the best of our knowledge, there are no applications in the literature that combine area-level SAE models and RFs using alternative data sources such as remote sensing data. The newly introduced area-level MERF is applied to estimate household consumption at the km grid-level in Mozambique based on survey and geospatial data.
Section 4.2 explains the statistical methodology: While Section 4.2.1 proposes an area-level MERF, a nonparametric bootstrap MSE estimator is introduced as a measure of uncertainty in Section 4.2.2. Model-based simulations are performed in Section 4.3 to evaluate the introduced methods. Section 4.4 describes the application of the area-level MERF for the estimation of consumption on a grid-level in Mozambique. Section 4.5 concludes and gives an outlook.

4.2 Model estimation

In the following, we present the statistical methodology in which we combine an area-level model with a RF. As usual, a direct estimator is computed with survey data and combined with a synthetic part that uses covariate information from additional data sources. For the synthetic part, we use a RF instead of a linear model. Before we start with the proposed model, a few remarks about RFs follow. RFs were first introduced by Breiman (2001) and are based on the concept of bootstrap aggregation (bagging) (Breiman, 1996). They can be used for regression or classification problems, focusing here only on regression. For estimation, a large number of bootstrap samples are drawn with which regression trees are computed. The predictions from the individual trees are averaged, which can reduce the noise, and thus the bias. By using regression trees, complex interactions between covariates and nonlinear relationships with the target variable can be detected. In addition, variance can be reduced by using a collection of de-correlated trees. This is achieved by randomly selecting a set of variables for splitting. Due to the aggregation of a large number of regression trees, RFs are outlier-robust. RFs cannot extrapolate because their predictions are based on the input data. This can be disadvantageous in applications with very different training and test data, but advantageous when the target variable is in a predefined range of values. For more details on RFs see Breiman (2001) and for an overview see Hastie et al. (2008).

4.2.1 Area-level mixed effects random forest

The general underlying model for area-level SAE models is a LMM. For clustered unit-level data, Hajjem et al. (2014) extend the concept of RFs to mixed effects RFs. In the context of SAE, Krennmair and Schmid (2022) broaden this idea to SAE applications with unit-level data. For area-level data, the methodology can be transferred as follows. A finite population of size N is assumed and partitioned into $d = 1, ..., D$ domains. A sample of size n with $i = 1, ..., n_d$ units per domain is drawn with domain-specific sample sizes n_d so that $n = \sum_{d=1}^{D} n_d$. Area-level models are typically divided into two stages: the sampling model and the linking model. The assumption of the sampling model is, that the direct estimator based on survey data can be represented by the domain-specific true indicator θ_d and a sampling error e_d:

$$\hat{\theta}_d^{Dir} = \theta_d + e_d, \quad e_d \overset{ind}{\sim} N(0, \sigma_{e_d}^2).$$

If the indicator of interest is a mean value, a common direct estimator is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952):

$$\hat{\theta}_d^{Dir} = \frac{\sum_{i=1}^{n_d} w_{di} y_{di}}{\sum_{i=1}^{n_d} w_{di}}, \tag{4.1}$$

where the w_{di} denote the sampling weights (inverse inclusion probabilities). The variance of the direct estimator $\sigma^2_{e_d}$ is assumed to be known, but usually has to be estimated from the unit-level sample data using either bootstrap methods (Alfons and Templ, 2013) or for example the Horvitz-Thompson approximation (Horvitz and Thompson, 1952). The linking model relates the covariate information from additional (population) data sources to the true indicator:

$$\theta_d = f(x_d) + v_d, \quad v_d \overset{iid}{\sim} N(0, \sigma^2_v), \tag{4.2}$$

where $f()$ denotes a RF and v_d domain-specific random effects with variance σ^2_v. The domain-specific covariate information is denoted by a $p \times 1$ vector x_d. The combination of both models leads to:

$$\hat{\theta}^{Dir}_d = \theta_d + e_d = f(x_d) + v_d + e_d, \quad v_d \overset{iid}{\sim} N(0, \sigma^2_v), \quad e_d \overset{ind}{\sim} N(0, \sigma^2_{e_d}). \tag{4.3}$$

The unknown components/parameters of the model are the random effects variance σ^2_v and the RF $f()$. Once these parameters are estimated, the final domain-specific estimator can be obtained as follows:

$$\begin{aligned} \hat{\theta}_d &= \hat{f}(x_d) + \hat{v}_d = \hat{\gamma}_d \hat{\theta}^{Dir}_d + (1 - \hat{\gamma}_d)\hat{f}(x_d), \quad \text{with} \\ \hat{v}_d &= \frac{\hat{\sigma}^2_v}{\sigma^2_{e_d} + \hat{\sigma}^2_v}[\hat{\theta}^{Dir}_d - \hat{f}(x_d)] \quad \text{and} \quad \hat{\gamma}_d = \frac{\hat{\sigma}^2_v}{\sigma^2_{e_d} + \hat{\sigma}^2_v}. \end{aligned} \tag{4.4}$$

The shrinkage factor $\hat{\gamma}_d$ has the same interpretation as in the standard FH model: the smaller the model variance $\hat{\sigma}^2_v$ is compared to the total variance, which includes the direct variances, the more weight is given to the synthetic component, and vice versa. Note that if $f(x_d)$ is replaced by $x^T_d\beta$, the models in Equations (4.2) and (4.3) and the estimator in Equation (4.4) result in the standard area-level model and estimator introduced by Fay and Herriot (1979) (FH model/estimator). In the case of the FH model, there are several ways to estimate the model variance σ^2_v, such as maximum likelihood (ML) or restricted ML (REML). Avila-Valdez et al. (2020) present an expectation maximization (EM) algorithm for the FH model for ML and REML estimation of the model components. To estimate the model in Equation (4.3), we adapt the EM algorithm for ML estimation proposed by Avila-Valdez et al. (2020). The algorithm iteratively takes as correct the synthetic component estimated using an RF and the random effects. For the synthetic component, as in the MERF algorithm of Hajjem et al. (2014), the *out-of-bag* (OOB) predictions are used, i.e. the predictions of trees corresponding to the bootstrap sample without the respective observation (Breiman, 2001). When estimating the synthetic component of an area-level model, we want the direct estimates with higher reliability to have more influence. In the Fay-Herriot model, therefore, the variances of the direct estimators are included in the calculation of the regression coefficients in order to correctly represent the relationship between the covariates and the direct estimators. In order to avoid bias in the RF as well, it is constructed using the inverse sample variances $\sigma^2_{e_d}$ as case weights, and more reliable direct estimators are given a higher weight in the estimation by being selected with a higher probability in the bootstrap sample. The proposed area-level MERF algorithm is as follows:

1. Initialize $r = 0$, choose a starting value for $\hat{\sigma}^{2(0)}_v$ and estimate a RF $\hat{f}()^{(0)}$ with $y \sim X$ using the inverse sampling variances as weights $\sigma^{-2}_{e_1}, .., \sigma^{-2}_{e_D}$ in the estimation. y is a vector with the direct estimators $y = (\hat{\theta}^{Dir}_1, ..., \hat{\theta}^{Dir}_D)$ and X a $D \times p$ matrix with the covariate information. Get $\hat{f}(X)^{(0)}_{OOB}$.
2. Set $r = r + 1$. Update the random effects $\hat{v}^{(r)} = (\hat{v}^{(r)}_1, ..., \hat{v}^{(r)}_D)$, the RF $\hat{f}(X)^{(r)}$ and $\hat{\sigma}^{2(r)}_v$:
   (a) Calculate

$$\hat{v}^{(r)} = \left(\frac{1}{\hat{\sigma}^{2(r-1)}} I_D + R^{-1}\right)^{-1} R^{-1}(y - \hat{f}(X)^{(r-1)}_{OOB}),$$

where $R = diag(\sigma^2_{e_1}, .., \sigma^2_{e_D})$ and I_D is the $D \times D$ identity matrix.

(b) Estimate a RF $\hat{f}()^{(r)}$ with dependent variable $y - \hat{v}^{(r)}$, covariates X and weights $\sigma^{-2}_{e_1}, .., \sigma^{-2}_{e_D}$ and get $\hat{f}(X)^{(r)}_{OOB}$.

(c) Calculate:

$$\hat{\sigma}^{2(r)}_v = \frac{1}{D}\left(\hat{v}'^{(r)}\hat{v}^{(r)} + tr\left[\left(\frac{1}{\hat{\sigma}^{2(r-1)}_v} I_D + R^{-1}\right)^{-1}\right]\right).$$

3. Repeat step 2 until convergence is reached.

4. Get $\hat{\sigma}^2_v$ and a final model/RF $\hat{f}()$.

In the EM algorithm, convergence is achieved when the log-likelihood can no longer be maximized. Therefore, the relative change of the log-likelihood $\frac{|l^{(r)} - l^{(r-1)}|}{l^{(r-1)}}$ is used as convergence criterion where $l = l(\sigma^2_v, f|\hat{\theta}^{Dir}_d) = -0.5 \sum^D_{d=1} log[2\pi(\sigma^2_v + \sigma^2_{e_d})] + [\hat{\theta}^{Dir}_d - f(x_d)]^2(\sigma^2_v + \sigma^2_{e_d})^{-1}$. Convergence is achieved when the relative change is below a certain threshold, such as 10^{-5}. Once the final model is estimated, the resulting area-level MERF estimator $\hat{\theta}^{MERF}_d$ can be calculated according to Equation (4.4). For domains that are not in the sample, so-called out-of-sample domains, the predictions from the final RF are used.

4.2.2 Uncertainty estimation

In order to assess the quality of the point estimates, it is essential to determine a measure of uncertainty. A common measure of reliability in the SAE literature is the MSE (Rao and Molina, 2015). In the case of an area-level LMM, probably the best known analytical estimator of the MSE is that of Prasad and Rao (1990), which is approximately unbiased. However, resampling techniques also compete with analytical estimators (Gonzalez-Manteiga et al., 2008a). Although Giusti et al. (2012) derive an analytical MSE estimator based on the results of Opsomer et al. (2008), they also provide a nonparametric bootstrap estimator for the MSE that outperforms the analytical one in their simulation. Another advantage of nonparametric bootstrap MSE estimators is that distributional assumptions can be avoided. Especially for more complex models, such as the semiparametric FH model (Giusti et al., 2012), estimators based on bootstrapping are an established alternative. In the context of tree-based methods methods Krennmair and Schmid (2022) propose a bootstrap MSE estimator in the unit-level mixed-effects RF. We follow these approaches and propose the following nonparametric bootstrap to estimate the MSE of $\hat{\theta}^{MERF}_d$:

1. Estimate $\hat{\sigma}^2_v$ and $\hat{f}()$ with the algorithm proposed in Section 4.2.1. Calculate $\hat{v}_d = \frac{\hat{\sigma}^2_v}{\sigma^2_{e_d} + \hat{\sigma}^2_v}[\hat{\theta}^{Dir}_d - \hat{f}(x_d)]$ using the direct estimator $\hat{\theta}^{Dir}_d$ and $\sigma^2_{e_d}$ for $d = 1, ..., D$.

2. Center and rescale the random effects $\hat{v}_d$ for $d = 1, ..., D$:

$$\hat{v}^{cs}_d = \frac{(\hat{v}_d - \frac{1}{D}\sum^D_{d=1} \hat{v}_d)\hat{\sigma}_v}{\sqrt{\frac{1}{D}\sum^D_{d=1}(\hat{v}_d - \frac{1}{D}\sum^D_{d=1} \hat{v}_d)^2}}$$

3. Center and rescale the sampling errors $\hat{e}_d$ for $d = 1, ..., D$:

$$\hat{e}_d = \hat{\theta}^{Dir}_d - \hat{f}(x_d) - \hat{v}_d$$
$$\hat{e}^{cs}_d = \frac{(\hat{e}_d - \frac{1}{D}\sum^D_{d=1} \hat{e}_d)\sigma_{e_d}}{\sqrt{\frac{1}{D}\sum^D_{d=1}(\hat{e}_d - \frac{1}{D}\sum^D_{d=1} \hat{e}_d)^2}}$$

4. For $b = 1, ..., B$:

(a) Draw a simple random sample with replacement of size D from $(\hat{v}_1^{cs}, ..., \hat{v}_D^{cs})$ to get $v_d^{(b)}$.

(b) Draw a simple random sample with replacement of size D from $(\hat{e}_1^{cs}, ..., \hat{e}_D^{cs})$ to get $e_d^{(b)}$.

(c) Calculate the true bootstrap population parameter: $\theta_d^{(b)} = \hat{f}(x_d) + v_d^{(b)}$.

(d) Simulate the bootstrap sample: $\hat{\theta}_d^{Dir(b)} = \hat{f}(x_d) + v_d^{(b)} + e_d^{(b)}$.

(e) Estimate $\hat{\sigma}_v^{2(b)}$ and $\hat{f}()^{(b)}$ using the bootstrap sample from the previous step with the algorithm proposed in Section 4.2.1.

(f) Estimate the bootstrap estimator $\hat{\theta}_d^{MERF(b)}$ with Equation (4.4).

5. Estimate the MSE:

$$\widehat{\text{MSE}}(\hat{\theta}_d^{MERF}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_d^{MERF(b)} - \theta_d^{(b)}\right)^2.$$

4.3 Simulation experiment

This section presents a simulation study to empirically evaluate the performance of the estimators proposed in Section 4.2. We aim to investigate the performance of the point estimator in terms of bias and efficiency. The MSE estimator proposed in Section 4.2.2 is evaluated with respect to bias. In area-level SAE applications, the number of observations is determined by the number of domains. In many examples, if the observation unit is an administrative area, only 100 to 300 domains may exist or be represented in the sample. A larger number of domains from 500 to 1000 is rarely found at administrative level. However, if no administratively defined domains are considered, but rather geographical ones, for example, defined by km grids, the number of domains can also be far above 1000. This is the case, for example, in the application in Section 4.4 where we have 1170 km grids in Mozambique available. In order to reflect this varying number of domains in the simulation, the scenarios are considered for different numbers of domains D, where $D \in \{200, 500, 1000, 2000\}$. We look at the following scenarios/models:

linear:	$y = 10 + 2x_1 - 2x_2$,
interaction:	$y = 10 + 2x_1x_2 - 2x_2^2$,
interaction noise:	$y = 10 + 2x_1x_2 - 2x_2^2$ with additional noise $z_1, ..., z_8 \sim U(0, 1)$.

In each scenario, the true parameter of interest is generated for each domain $d = 1, .., D$ by $\theta = y + v$, with $v \sim N(0, 0.04)$. The direct estimates are then constructed as $\hat{\theta}^{Dir} = \theta + e$ with $e \sim N(0, \sigma_e^2)$. σ_e^2 is chosen in the same range as in Giusti et al. (2012). To ensure variation in sampling variances at large D, we choose $\sigma_e^2 \sim U(0.08, 0.16)$ and keep it fixed when generating the Monte Carlo replications per setting. The auxiliary variables x_1 and x_2 are distributed $U(0, 1)$. We study the behavior of the proposed estimator under different relationships between the dependent variable and the auxiliary variable. Similar to Krennmair and Schmid (2022), we consider a linear model (*linear*) and a model with a nonlinear relationship and interaction (*interaction*). In addition, we examine the same interaction scenario with additional noise variables to see how the estimator performs with required automatic variable selection (*interaction noise*). For each scenario and number of domains D, the data is generated $M = 500$ times. For point estimation, we consider two competing estimators: the FH estimator (Fay and Herriot, 1979) and a semiparametric FH (SPFH) estimator proposed by Giusti et al. (2012). For the implementation of the simulation, R (R Core Team, 2022) is used. The FH estimator is computed using the R-package **emdi** (Kreutzmann et al., 2019), while the SPFH estimator is calculated using code provided by the authors of Giusti et al. (2012). The RF of the proposed area-level MERF estimator is estimated using the R-package **ranger** (Wright and Ziegler, 2017) with default settings. We expect that in the *linear* setting the FH estimator performs at least as well as the MERF, or even

better. The effect should be seen especially with a small number of domains. When the number of observations is small, the variation in the bootstrap samples may not be sufficient to detect complex relationships. In the *linear* setting, the relationship is not complex, but corresponds exactly to the linear structure of the FH model. This is also true for the SPFH, since a linear function is a special case of a spline function. In the *interaction* scenario, the MERF should have an advantage over the FH estimator because the interaction terms cannot be accounted for by the FH estimator. This effect should increase as the number of domains increases. Compared to the SPFH, the MERF should be at least as efficient. The *interaction* scenario contains an interaction as well as a quadratic term that is accounted for by the penalized splines of the SPFH, while the interaction is not accounted for. The MERF should account for both the interaction and the quadratic term. In the interaction scenario with additional noise variables (*interaction noise*), variable selection is performed in advance for the FH estimator for each Monte Carlo sample with the explanatory and noise variables using the Akaike information criterion (AIC). The FH estimator is then estimated using only the resulting variables. The same set of variables is then used for the SPFH estimator. Accordingly, fewer variables enter the FH and SPFH estimators in the model estimation in the *interaction noise* scenario than in the MERF approach. As the MERF automatically detects the relevant variables, both the explanatory and noise variables are passed. Therefore, we expect the MERF to outperform the other two comparative estimators in this scenario, at least with a large number of domains. The quality of the point estimators is evaluated using the relative bias (RB) and the empirical root MSE (RMSE) defined in Equation (4.5). To judge the proposed bootstrap estimator in terms of bias, the relative bias of the RMSE (RB RMSE), defined in Equation (4.6) is used.

$$\text{RB}(\hat{\theta}_d) = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{\theta}_{d_r} - \theta_{d_r}}{\theta_{d_r}}\right), \qquad \text{RMSE}(\hat{\theta}_d) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{d_r} - \theta_{d_r}\right)^2}, \tag{4.5}$$

$$\text{RB RMSE}(\hat{\theta}_d) = \frac{\frac{1}{R}\sum_{m=1}^{M}\sqrt{\widehat{\text{MSE}}_{d_r}} - \text{RMSE}(\hat{\theta}_d)}{\text{RMSE}(\hat{\theta}_d)}. \tag{4.6}$$

Examining the results of the simulation experiment, we first look at the point estimators. The mean and median values of the RB and RMSE are shown per scenario and domain size for the three estimators in Table 4.1. First, we note that the MERF is unbiased in each scenario and at each domain size D. Nevertheless, the values of the RB are higher than for the FH and SPFH estimators. This difference decreases with higher domain sizes and even reverses for the *interaction* scenario at least for 1000 observations/domains. We now turn to the efficiency of the MERF approach compared to the two benchmark estimators. Figure 4.1 shows boxplots of the distribution of RMSEs for the estimators in each setting. In the *linear* setting, there is a clear efficiency advantage of the FH and SPFH estimator over the MERF. The advantage diminishes a little with higher domain sizes, but does not disappear. As expected, the FH estimator is the least efficient in the *interaction* scenario because the underlying data generation process does not match the model. SPFH and MERF perform equally well at $D = 200$. At higher domain sizes, the efficiency of the MERF overcomes that of the SPFH because the MERF accounts for both the interaction and the quadratic term. At first glance, the results of the *interaction noise* scenario are somewhat counterintuitive. However, considering that FH and SPFH are already estimated with the covariates following a variable selection procedure so that the uncertainty of finding the true relationship with the dependent variable is not taken into account in the estimation, the results seem plausible for small domain sizes. We see that for $D = 500$ the MERF already performs better than the FH estimator and almost on par with the SPFH. For $D = 1000$ and $D = 2000$, the MERF approach is more efficient. We do not compare the FH and SPFH estimators with the full set of variables (explanatory variables and noise) in the *interaction noise* scenario, as this would not correspond to common practice in applications and the comparison would not be fair. Table 4.2 helps understanding the results of Figure 4.1, especially for the *linear* scenario. Table 4.2 contains averages over the simulation runs of the R^2 and the estimated

Table 4.1: Mean and Median of RB[%] and RMSE for varying domain sizes D.

	D	200		500		1000		2000	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
RB									
linear	FH	0.040	0.041	0.018	0.015	0.031	0.031	0.032	0.032
	SPFH	0.044	0.042	0.019	0.015	0.032	0.034	0.033	0.032
	MERF	0.089	0.091	0.043	0.046	0.046	0.047	0.040	0.037
interaction	FH	0.056	0.051	0.051	0.055	0.053	0.052	0.062	0.062
	SPFH	0.048	0.046	0.043	0.041	0.045	0.042	0.053	0.054
	MERF	0.079	0.076	0.052	0.049	0.042	0.041	0.042	0.044
interaction noise	FH	0.062	0.062	0.063	0.060	0.056	0.058	0.052	0.050
	SPFH	0.056	0.059	0.055	0.050	0.048	0.047	0.043	0.041
	MERF	0.091	0.091	0.083	0.083	0.070	0.074	0.057	0.054
RMSE									
linear	FH	0.179	0.179	0.175	0.176	0.174	0.174	0.173	0.173
	SPFH	0.185	0.185	0.178	0.179	0.175	0.176	0.174	0.174
	MERF	0.217	0.217	0.204	0.204	0.199	0.199	0.197	0.197
interaction	FH	0.229	0.230	0.227	0.228	0.226	0.227	0.226	0.227
	SPFH	0.214	0.215	0.210	0.211	0.209	0.209	0.208	0.209
	MERF	0.209	0.209	0.201	0.201	0.197	0.197	0.195	0.195
interaction noise	FH	0.232	0.232	0.229	0.230	0.227	0.227	0.226	0.227
	SPFH	0.219	0.220	0.213	0.213	0.210	0.210	0.209	0.209
	MERF	0.237	0.237	0.215	0.215	0.202	0.203	0.193	0.193

model variance $\hat{\sigma}_v^2$ for the FH model and the MERF approach. For the FH model, the R^2 specifically applicable to FH models (Lahiri and Suntornchost, 2015) is used and for the MERF we look at the OOB R^2 of the final random forest from the proposed algorithm in Section 4.2.1. Although the two measures are not directly comparable, since the former evaluates the goodness of the entire model, and the latter only the structural part, they help to understand the results as follows. For the MERF approach, the following relationship emerges for each scenario: As the domain size D increases, the R^2 increases, which means that the explanatory power of the random forest grows. Since it is an iterative algorithm that moves from forest building to estimating the variance of the random effects, the estimated variance of the random effects converges to the true model variance of $\sigma_v^2 = 0.04$. Comparing the MERF values to the FH values, we find that the estimated model variances of the FH are smaller than those estimated using the MERF approach when the explanatory power of the FH model is higher, especially in the *linear* scenario. Smaller model variances mean that more weight is given to the synthetic component, leading to an increase in efficiency. In addition to these explanations, it is worth noting that the MERF approach is not performing poorly in the *linear* setting. The explanatory power is even greater than in the *interaction* setting, and the model variance is also very well estimated. The MERF approach is simply not better than the less complex linear model in the simplest scenario. We proceed with the investigation of the performance of the proposed MSE estimator resulting from the suggested nonparametric bootstrap procedure. In each simulation run, the number of bootstrap replications is set to $B = 200$. Table 4.3 contains mean and median values across domains for each setting. Generally, we observe that the proposed MSE estimator suffers from a slight overestimation. There are some differences between the number of domains and the scenarios, and interestingly, the overestimation becomes smaller as the complexity of the scenario increases. Due to the slight overestimation, the proposed MSE estimator can be considered somewhat conservative. Nevertheless, all mean and median values are below 10% and

can therefore be seen as a reasonable approximation to the uncertainty of the proposed MERF approach. Detailed results on the RB RMSE can be found in the Appendix in Figure D.1.

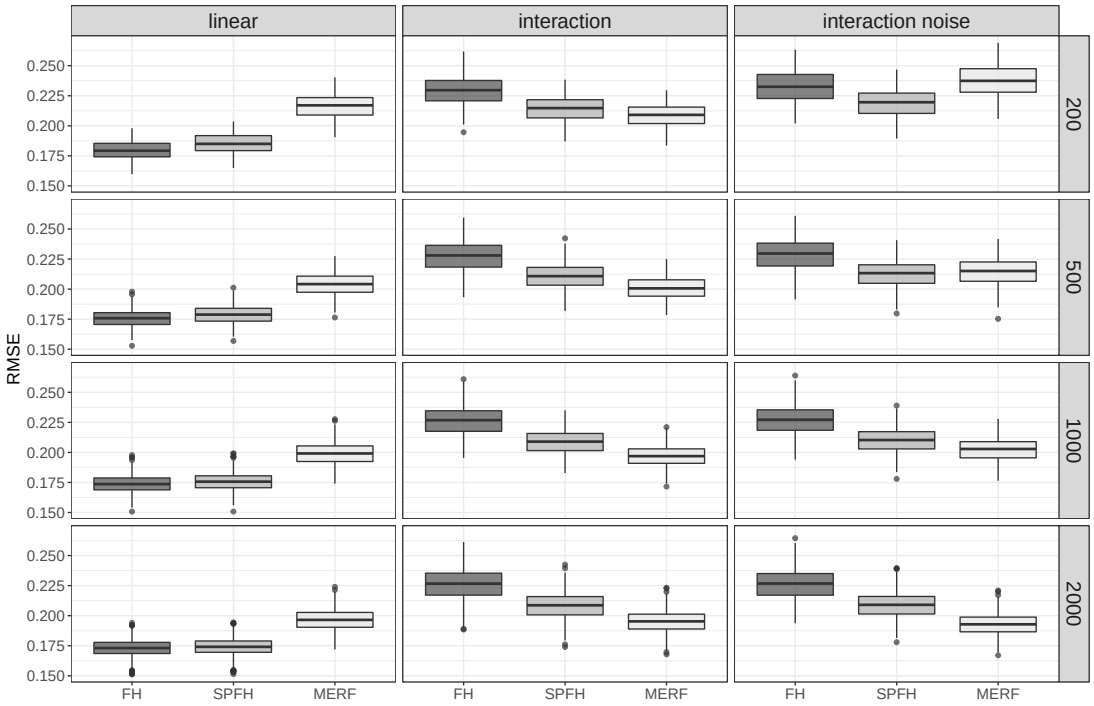


Figure 4.1: Distribution of RMSEs for varying domain sizes D.

Table 4.2: Averages of R^2 and $\hat{\sigma}_v^2$ over $M = 500$ simulation runs.

D		200	500	1000	2000
R^2					
linear	FH	0.917	0.917	0.917	0.917
	MERF	0.883	0.882	0.883	0.884
interaction	FH	0.609	0.606	0.607	0.610
	MERF	0.698	0.702	0.707	0.710
interaction noise	FH	0.621	0.617	0.611	0.609
	MERF	0.635	0.669	0.683	0.692
$\hat{\sigma}_v^2$					
linear	FH	0.038	0.039	0.040	0.040
	MERF	0.060	0.051	0.049	0.048
interaction	FH	0.086	0.088	0.089	0.089
	MERF	0.054	0.049	0.047	0.046
interaction noise	FH	0.082	0.086	0.088	0.089
	MERF	0.093	0.069	0.058	0.052

Table 4.3: Mean and Median of RB RMSE[%].

D	200		500		1000		2000	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
linear	6.139	5.919	7.091	7.300	7.675	7.664	7.274	7.115
interaction	6.024	5.852	6.884	6.920	7.255	7.184	7.134	7.148
interaction noise	2.197	2.302	4.107	3.982	4.379	4.283	5.058	4.993

4.4 Application

The aim of this section is to illustrate the area-level MERF presented in Section 4.2 and to investigate its performance using a real world example. For this, we estimate the consumption per capita for Mozambique on a 1km square grid-level as SAE area-level. Mozambique ranks 185 out of 191 countries in the United Nations Development Program's latest Human Development Index, which takes into account gross national income, life expectancy and access to education (United Nations Development Programme, 2022), making it one of the poorest countries in the world. While Mozambique's per capita growth rates were increasing between 2001 and 2015, the growth rate turned negative since 2016 due to a hidden debt crisis, cyclones that mainly hit Northern and Central provinces in 2019 and the COVID-19 pandemic in 2020. A positive trend has been recorded again since 2021 (Da Maia, 2022). But even in the times of economic growth, poverty levels differ significantly across the country. The differences mainly persist between rural and urban regions. Southern provinces are comparatively more wealthy than Northern and Central provinces which can be partly explained by the higher degree of urbanization in the South. The capital Maputo also lies in the South of Mozambique.

4.4.1 Data

We use traditional survey data from a household welfare survey and, as auxiliary information, geospatial data. Both are from 2019 and were kindly provided by the World Bank Group. Mozambique is divided into 11 provinces. Due to a lack of reliable data, the province Niassa is not considered for this exemplary application. The remaining 10 provinces are Cabo Delgado, Gaza, Inhambane, Manica, Maputo City, Maputo, Nampula, Sofala, Tete and Zambezia. As variable of interest serves the average household consumption per capita in Mozambican Metical (MZN), which is spatially deflated using estimated local prices. We obtain the direct estimates by computing the Horvitz-Thompson estimator defined in Equation (4.1) per grid of household-level survey data. Following Edochie et al. (2023), the variance of the direct estimator is estimated by using the Horvitz-Thompson approximation of the R-package **sae** (Molina and Marhuenda, 2015). Therefore, the sum of sample weights per grid is calculated as approximation of the domain size. The results of the direct estimator are presented in the next section. Of the 41137 available grids, 1170 are in-sample and 39967 are out-of-sample grids. A summary of the sample sizes over grids is given in Table 4.4. The sample sizes of the in-sample grids range from 3 to 24 with a mean of 10.3.

Table 4.4: Distribution of sample sizes for grids

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Sample size	3.00	9.00	9.00	10.30	12.00	24.00

The auxiliary information stems from satellite data and includes in total 37 covariates that contain different information about buildings, land coverage, night light and rainfall. Summaries of the distribution of the possible auxiliary information are provided in Table D.1 in the Appendix. Since several variables for the same type of measure, for example minimum, mean and maximum rainfall, are included, potential

interactions between the covariates influence the model estimation.

4.4.2 Model estimation and results

The point and MSE estimates of the MERF are estimated as proposed in Section 4.2. For the RF part of the model, the `ranger` function of the R-package **ranger** (Wright and Ziegler, 2017) is utilized with default settings. The convergence criterion of the EM algorithm is set to 10^{-5}. For the nonparametric bootstrap MSE estimation, $B = 200$ bootstrap replications are performed. Common measures for the interpretation of RF type models are variance importance plots and partial dependence plots for influential covariates (Greenwell, 2017) (see Figure D.2 in the Appendix). For some of the covariates (i.a. the land cover variables lc_shrubcoverf, lc_grasscoverf, lc_cropscoverf), nonlinear relations to the dependent variable are recognizable. The MERF results are not only compared to the direct estimates, but also to a traditional FH model. The normality assumptions of a standard FH model were not fulfilled, thus we apply a log transformation. For the estimation of the FH model, the `fh` function of the R-package **emdi** (Kreutzmann et al., 2019) is used with ML variance estimation, log transformation and crude backtransformation which is the suggested backtransformation of package **emdi** in the presence of out-of-sample domains. An analytical MSE estimator for the crude backtransformation following Rao and Molina (2015) and Datta and Lahiri (2000) is automatically provided by **emdi**. For the variable selection of the FH model, we made use of the AIC based on a linear regression model with a log transformed target variable. We proceed with the presentation of the results. Table 4.5 contains the

Table 4.5: Distribution of direct and model-based point estimates, CVs [%] and shrinkage factors for model-based estimates.

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
In-sample							
Point estimates	Direct	1.01	28.79	43.83	66.22	75.56	2023.41
	log FH	14.27	30.72	42.89	58.89	68.14	593.10
	MERF	1.01	30.15	38.31	44.73	54.96	112.68
CV	Direct	23.63	34.26	36.20	37.61	39.25	81.03
	log FH	18.70	22.88	23.44	23.70	24.26	46.59
	MERF	9.79	16.10	19.51	21.31	23.22	81.86
Shrinkage factors	log FH	0.12	0.36	0.39	0.38	0.42	0.60
	MERF	0.00	0.09	0.25	0.29	0.44	1.00
Out-of-sample							
Point estimates	log FH	0.10	27.70	31.41	35.48	37.65	722.12
	MERF	18.70	27.88	30.61	32.45	34.51	117.14
CV	log FH	28.57	28.77	28.86	29.11	29.02	186.66
	MERF	9.65	27.35	30.58	30.58	33.95	56.62

distribution of direct and model-based point and coefficient of variation (CV) estimates, as well as model-based shrinkage factors. Additionally, boxplots of the point estimates for in- and out-of-sample grids are provided in the Appendix (Figure D.3). The direct point estimates range from 1.01 to 2023.41 MZN, indicating a right-skewed distribution with differing median and mean values of 43.83 and 66.22 MZN, respectively. The CVs of the direct estimator vary between 23.63 and 81.03 with a mean of 37.61 and are therefore far above a threshold of 20% that is traditionally considered reliable (Eurostat, 2023). The median values of the point estimates of the direct and model-based estimates are very similar, only the MERF estimates are slightly lower. The interquartile ranges (1st Qu. and 3rd Qu.) show less variation in both SAE models, MERF and log FH, than in the direct estimates. This is a result of shrinkage in both SAE models, where unreliable direct estimates are shrunk to the center due to very

small sample sizes at the grid-level (see Table 4.4). Another point that becomes clear is illustrated in Figure 4.2, which shows line plots of the (in-sample) point estimates for four provinces in Mozambique as an example. Again, we see that the model-based estimates generally follow the direct estimates, but with less variation. In addition, we see that the MERF smooths more than log FH, and high peaks are not tracked. This can be explained by the robustness to outliers property of RFs (Breiman, 2001), but might be too extreme in some situations in the area-level context. For example, if the distribution of direct estimates is highly skewed, but the point estimates are more or less reliable. On the other hand, if the direct estimates are not reliable, the outlier robustness of the MERF may be beneficial and the stronger smoothing may be appropriate. Turning to the uncertainty of the point and model-based

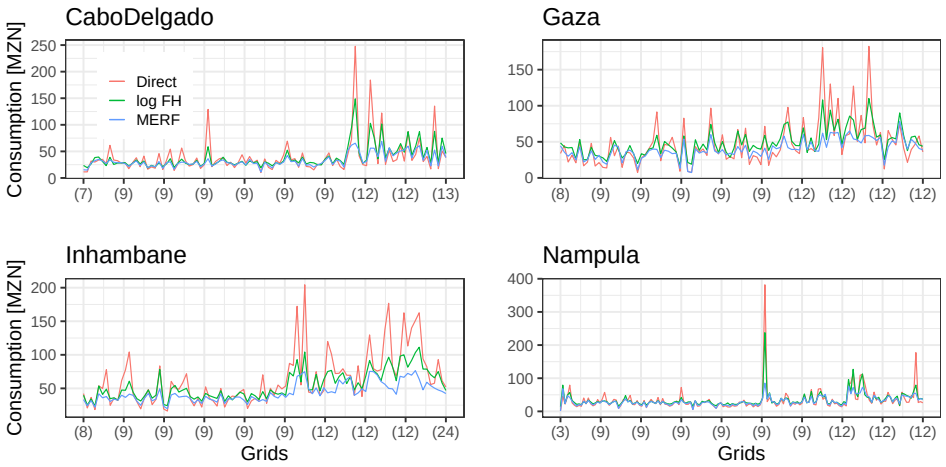


Figure 4.2: Line plots of the in-sample point estimates for the Provinces Cabo Delgado, Gaza, Inhambane and Nampula. The grids are ordered by increasing sample size, with the sample size of every 10th grid in parentheses.

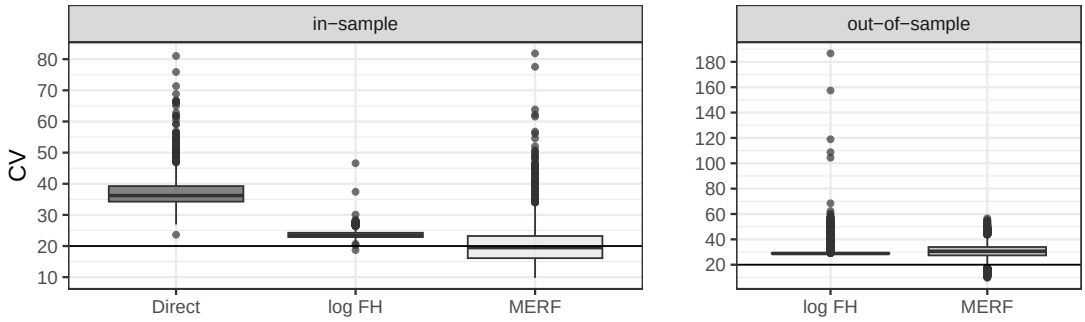


Figure 4.3: Boxplots of CVs [%] separated for in- and out-of-sample grids.

estimates, Figure 4.3 shows boxplots of the CVs for in- and out-of-sample grids. It is striking that the use of both models significantly improves the reliability of the results compared to direct estimates, as measured by the CV. For in-sample grids, the reduction of the CVs is the largest for the MERF, followed by log FH. The proposed MERF is more efficient here than log FH. One reason for this lies in the fact, that the shrinkage factors of the MERF are on average smaller than those of log FH. The median values of the shrinkage factors are for example 0.25 and 0.38 for the MERF and log FH, respectively (Table 4.5). Smaller shrinkage factors indicate that more weight is put on the synthetic part of the model, in this case the RF part, and less on the direct estimator with its higher variances. For out-of-sample grids, the median of the MERF is slightly larger than that of log FH (30.58 vs. 28.86, see Table 4.5),

but log FH reaches much higher outlier and maximum values than the MERF. The maximum of the MERF amounts to 56.62, while of log FH is 186.66. Finally, we discuss the actual results of the model

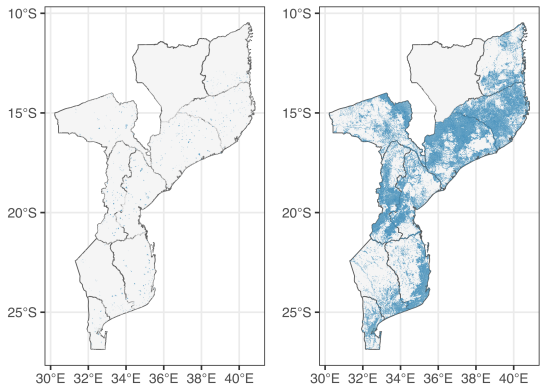


Figure 4.4: Available grids for the direct (left) and MERF (right) estimates for Mozambique.

estimates for Mozambique and focus on the newly proposed MERF approach. Figure 4.4 shows a great advantage of using SAE models. The grids where the direct and MERF estimates are available for Mozambique are plotted. Without using any model-based method, it is barely possible to draw any conclusions on such a disaggregated level like the grid-level (left map). With the help of the MERF approach, predictions for almost the whole country are provided (right map). The gray region at the top of the map belongs to the province of Niassa which is excluded from the analysis. Satellite data was not available for the remaining gray areas, which are likely uninhabited. For an easier comparison, a geographic map of Mozambique is provided in the Appendix (Figure D.4). Figure 4.5 plots the MERF

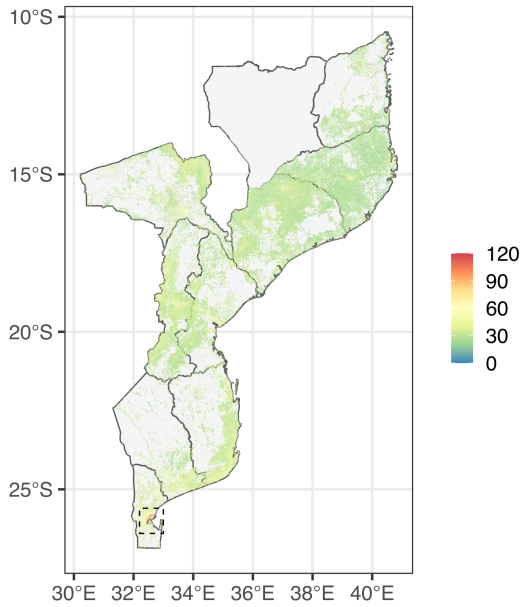
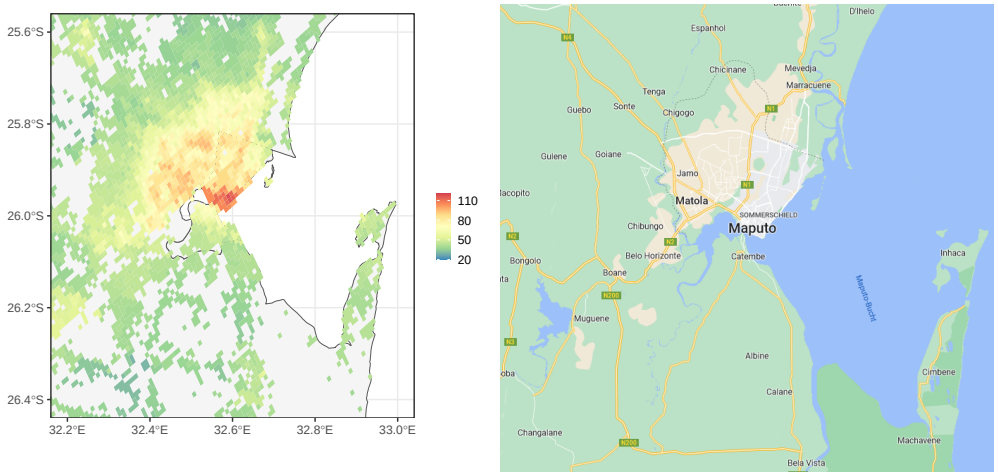


Figure 4.5: MERF estimates of consumption [MZN] on a grid-level for Mozambique.

predictions of the average consumption per grid to help identify regional differences. While green and yellow grids dominate the map indicating consumptions levels ranging from 30 to 55 MZN, also wealthier regions are recognizable. The southern region around the two largest cities of Matola and Maputo (small dotted box), the third largest city of Nampula (15° S, 39° E), and the city of Beira (19.8° S, 35° E) are characterized by consumption levels of around 85 to 120 MZN. This finding corresponds to the poverty situation in Mozambique described at the beginning of this section. The country as a

whole is one of the poorest in the world, but there are large differences between urban and rural areas and between northern and southern provinces (Santos and Salvucci, 2016). Finally, Figure 4.6 provides a closer look at the region around the capital Maputo: a) MERF estimates of consumption [MZN] for the provinces Maputo and Maputo City and b) the corresponding Google Maps extract. The greater Maputo Area is the prime urban agglomeration in Mozambique. Especially the southern neighborhoods of Sommershield in Maputo report the highest levels of consumption. This finding is in line with the World Bank Working Paper of Herzog et al. (2017) about urban poverty in the Greater Maputo Area. In



(a) MERF estimates (b) Google Maps extract

Figure 4.6: Results and map of the provinces Maputo and Maputo City.

conclusion, in the real data SAE example, the area-level MERF represents a valuable alternative to the traditional FH model, especially in terms of efficiency. Other advantages are that possible interactions between the covariates and nonlinear relationships of the covariates to the target variable are taken into account and that variables are selected automatically. A possible disadvantage is that the model smooths more than the log FH. Nevertheless, the MERF is able to produce reasonable results and captures the country-wide trends for consumption at the grid-level in Mozambique, which could also be confirmed by other sources.

4.5 Concluding remarks

The purpose of this paper is to propose a first way to combine area-level SAE models with RFs to allow for interactions, nonlinear relationships, and implicit variable selection. The results of the simulation experiment and application to real world data are encouraging and have potential for further research. In particular, the simulation results show that the proposed point estimator leads to unbiased estimates. In terms of efficiency, the results highlight that the presented approach can improve the efficiency with a large number of domains in the presence of interactions and additional noise variables that involve automatic variable selection compared to a linear and a spline-based estimator. However, the results also show the limitations of our methodology. For a rather small to medium number of domains (200 and 500) and in the case of linear relationships, classical SAE models such as the FH model are still convincing and preferable to RF-based approaches. However, RFs are also applicable when the number of covariates is too large or even exceeds the number of observations to estimate LMMs. The proposed bootstrap scheme for estimating the MSE of the point estimator is proven in simulation and leads to reliable uncertainty measures. An illustration of the methodology using aggregated household survey

and satellite data from Mozambique with km grids as the unit of observation shows that the approach leads to an improvement in efficiency compared to the direct estimator and also to the log-transformed FH estimator. One point that stands out in the application is that the MERF smooths more than, for example, the log-transformed FH estimator, due to the property of an RF to be robust to outliers. This property can be advantageous in the case of unreliable direct estimators, but possibly disadvantageous when the distribution of direct estimators is highly skewed, but the point estimators are mostly trusted. Therefore, investigating the use of transformations of the dependent variable to achieve a more symmetric distribution and to move extreme observations toward the center of the distribution may be of interest in the context of RF. In this paper the mean of an interval scaled variable is estimated. When it comes to estimating ratios such as the head-count ratio (Foster et al., 1984) or nonlinear indicators such as Gini coefficients (Gini, 1912), the Fay-Herriot model must be estimated using an appropriate transformation to ensure the correct range of values for the estimates. This usually requires the development of appropriate back-transformations and MSE estimators. Since the MERF cannot extrapolate, the transformation can be omitted here to achieve the desired range of values. It remains part of further research to investigate how well the MERF approach performs at the area-level in estimating other indicators such as head-count ratios or Gini coefficients. There is an ongoing debate as to whether area-level or unit-level models are preferable when combining household survey data and grid-level remote sensing data (Masaki et al., 2022; Newhouse et al., 2022). Aggregating the resulting grid-level estimates to a higher administrative level and comparing them with reliable direct or model-based estimators could therefore be of great value for research and other applications where survey and satellite data are available.

Appendix D

D.1 Appendix

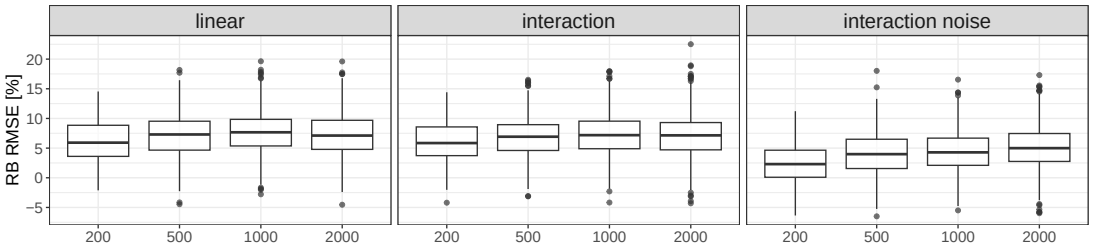


Figure D.1: Distribution of RB RMSEs [%] for varying domain sizes D (x-axis).

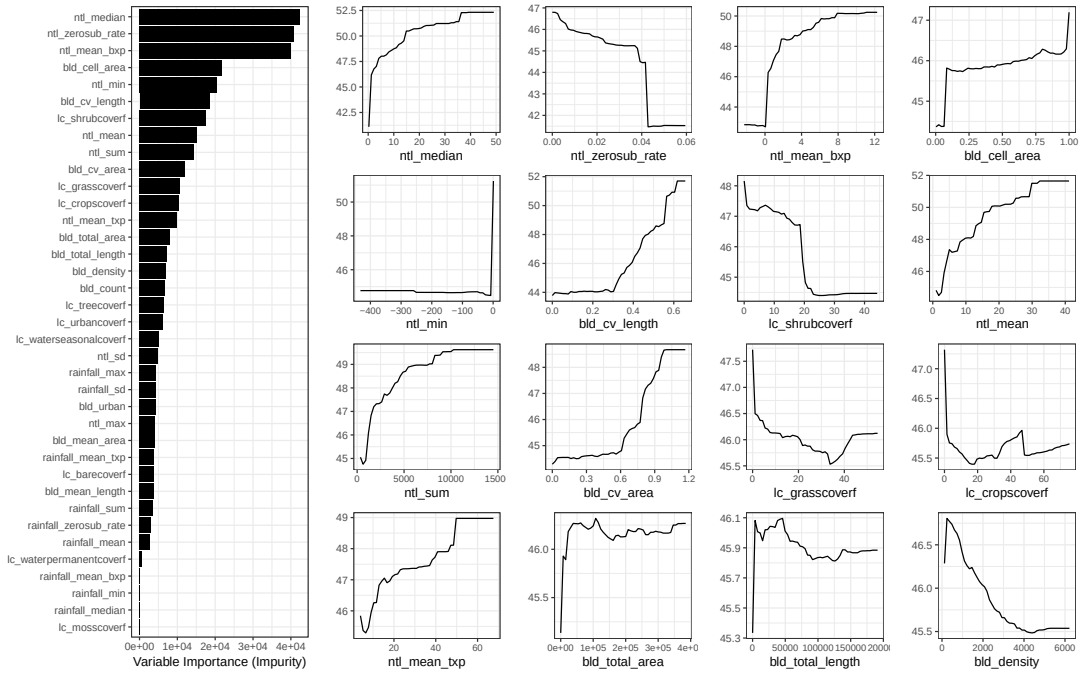


Figure D.2: Variable importance plot (left) and partial dependence plots (right) with consumption [MZN] on the y-axis of the 16 most influential variables.

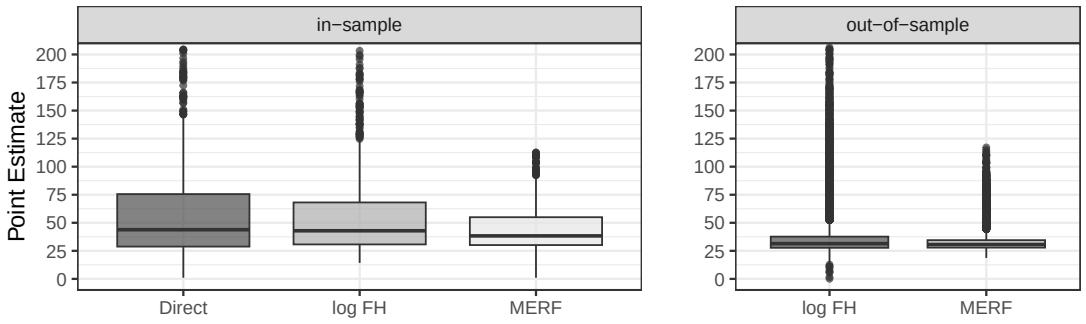


Figure D.3: Boxplots of point estimates separated for in- and out-of-sample grids. For an improved readability of the boxplots, some extreme values of the log FH model have been omitted. A summary of the whole distribution of log FH is provided in Table 4.5.

Table D.1: Distributions of possible geospatial auxiliary information on buildings (bld), land coverage (lc), night light (ntl) and rainfall.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
bld_cell_area	0.000	0.113	0.181	0.247	0.292	1.000
bld_count	0.367	36.667	66.601	206.132	146.871	6205.344
bld_cv_area	0.000	0.244	0.308	0.340	0.402	1.246
bld_cv_length	0.000	0.129	0.164	0.182	0.215	0.856
bld_density	119.007	270.582	350.441	543.176	589.357	6217.285
bld_mean_area	6.538	22.818	28.352	32.975	35.637	2682.253
bld_mean_length	10.419	18.696	20.748	21.599	23.199	192.342
bld_total_area	4.713	930.420	1829.879	9321.449	4425.757	384702.531
bld_total_length	5.525	729.769	1362.685	5246.629	3146.822	191125.125
bld_urban	0.000	0.000	0.000	0.086	0.000	1.000
lc_barecoverf	0.000	0.024	0.503	0.954	1.231	54.673
lc_cropscoverf	0.000	15.036	21.683	22.903	28.145	78.734
lc_grasscoverf	0.000	24.360	29.887	29.058	34.151	80.086
lc_mosscoverf	0.000	0.000	0.000	0.000	0.000	1.444
lc_shrubcoverf	0.000	17.977	21.914	20.869	24.783	47.188
lc_treecoverf	0.000	13.253	19.007	21.464	27.704	81.969
lc_urbancoverf	0.000	0.058	0.699	6.046	3.188	100.000
lc_waterpermanentcoverf	0.000	0.000	0.000	0.447	0.000	95.562
lc_waterseasonalcoverf	0.000	0.000	0.000	0.504	0.000	79.880
ntl_max	19.728	30.473	35.516	40.072	42.582	7996.105
ntl_mean	0.783	1.876	2.066	2.413	2.301	41.145
ntl_mean_bxp	-3.156	0.085	0.093	0.280	0.108	12.440
ntl_mean_txp	3.790	5.124	5.658	6.105	6.274	74.458
ntl_median	0.231	0.288	0.301	0.746	0.328	49.101
ntl_min	-445.389	-0.227	-0.181	-1.839	-0.135	0.880
ntl_sd	2.493	3.865	4.280	4.474	4.678	425.616
ntl_sum	276.485	662.232	729.379	851.842	812.105	14524.128
ntl_zerosub_rate	0.000	0.020	0.030	0.027	0.036	0.071
rainfall_max	-9999.001	50.841	62.785	-14.840	77.560	202.355
rainfall_mean	-9999.001	2.671	3.368	-77.701	3.952	6.332
rainfall_mean_bxp	-9999.001	0.000	0.000	-80.995	0.000	0.000
rainfall_mean_txp	-9999.001	2.671	3.369	-77.641	3.955	17.187
rainfall_median	-9999.001	0.000	0.000	-80.995	0.000	0.000
rainfall_min	-9999.001	0.000	0.000	-80.995	0.000	0.000
rainfall_sd	0.000	7.303	8.561	8.466	9.733	16.167
rainfall_sum	-3649635.356	974.924	1229.429	-28360.823	1442.582	2311.286
rainfall_zerosub_rate	0.644	0.723	0.754	0.761	0.800	1.000

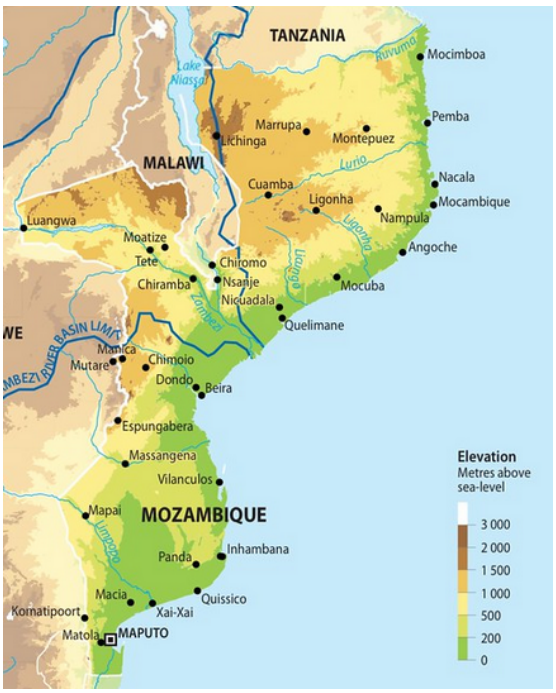


Figure D.4: Geographic map of Mozambique. *Source: GRID-Arendal, https://www.grida.no/resources/5160*

Chapter 5

Estimating intra-regional inequality with an application to German spatial planning regions

5.1 Motivation

For some time now, and especially since the United Nations Sustainable Development Goals (SDGs) of 2016, the reduction of inequality within and among countries has increasingly become a focus of public debate. Regionally differentiated indicators to measure poverty and inequality are thereby receiving growing attention in the attempt to quantify inequality. In order to meet the demands and expand policies to reduce economic inequality, it is of great importance to provide reliable statistics that adequately capture regional differences in income inequality. In Germany, due to its division in 1949 and reunification in 1990, economic inequality, especially between East and West, has been a particular focus of political and public debate. At the latest since the financial crisis of 2008/2009 regional income and wealth disparities that go far beyond East and West have reached public awareness, and this is likely to be reinforced with the 2020/2021 pandemic. Therefore, Goebel and Frick (2005) already considered regional income stratification by dividing Germany into four parts. Braml and Felbermayr (2018) focus on inequality at the county level measured by gross domestic product per capita, just as Kreutzmann et al. (2022) consider regional heterogeneity in wealth. In both, the focus is on the difference between regions, while an additional aspect of inequality is income differences between households within a region. Immel and Peichl (2020) combine both perspectives and look at regional income inequality at the county level measured by the top 10% earners and the bottom 40% within regions. When examining the regional dimension of income distributions, a distinction must be made between intra- and inter-regional inequality, as noted before. When considering intra-regional inequality, an appropriate measure must be used to determine the level of income inequality. A popular indicator for this purpose is the Gini coefficient (Gini, 1912), which is defined between zero and one, where zero means perfect equality and one maximum inequality. The presented methodology is illustrated by estimating Gini coefficients at

a regionally dis-aggregated level for Germany, which additionally represents to best of knowledge the first attempt to estimate Gini coefficients for Germany at a regional level lower than the federal states.

When it comes to measuring regional differences, the level of observation can become very detailed and the unit sample sizes very small. A unit in this context can be a regional area, a socio-demographically defined domain or a combination of both. In either case it is referred to as a domain or an area and when sample sizes are small, as small area. For small sample sizes, common estimators that use only survey data (hereafter referred to as direct estimators) are often not accurate enough to provide reliable domain-specific estimates of an indicator of interest. In these cases, small area estimation (SAE) methods allow for an increase in accuracy. In particular, model-based SAE methods use related additional data sources and information from other areas for this purpose. Overviews of SAE methods can be found in Pfeffermann (2013), Rao and Molina (2015) and Jiang and Rao (2020). A general framework for the construction of small area statistics is presented by Tzavidis et al. (2018). In Pratesi (2016) SAE methods particularly for the analysis of poverty data are provided. The most common SAE methods to estimate poverty and inequality indicators, such as Gini coefficients, on a dis-aggregated level are the World Bank method proposed by Elbers et al. (2003) or the empirical best predictor (EBP) method proposed by Molina and Rao (2010). In practice, however, this is problematic for privacy reasons. Especially when it comes to population data on a micro/individual level that are needed as auxiliary information. In these cases, area-level methods can help, where survey and related population data are only needed at the aggregated level. In addition, area-level models account for complex survey designs in the estimation of point and variance estimators. One of the most popular area-level SAE models is that proposed by Fay and Herriot (1979), known as the Fay-Herriot (FH) model, which is the underlying statistical model in this paper. In addition, there are empirical and hierarchical Bayesian methods, see for a comprehensive overview for example Rao and Molina (2015). In particular, the FH model can be estimated by a hierarchical Bayes model as well. Liu et al. (2014) use the hierarchical Bayes version of the FH model to compare it to a normal-logistic and a beta-logistic Bayes model for the use-case of estimating small area proportions. Also Janicki (2020) studies a hierarchical Bayesian model with a Beta distribution and a logit link to estimate poverty rates. The common property of proportions and Gini coefficients is that both are bounded in the interval $(0, 1)$. Therefore, some of the methods can be used for both applications. Fabrizi and Trivisano (2016) propose a hierarchical Beta mixed Bayesian regression area-level model with a logit link to estimate Gini coefficients for small areas and Fabrizi et al. (2016) apply this approach to jointly estimate at-risk-of-poverty rates and the Gini coefficients. The advantages of this and more general Bayesian approaches are that from the resulting posterior distribution, which is approximated by a Markov Chain Monte Carlo (MCMC) algorithm, the point estimates are directly given with an uncertainty measure as well as credible intervals. The possibility to specify different prior distributions of the model parameters also makes the model quite flexible. However, frequentist approaches probably predominate in the SAE literature and are widely accepted in National Statistical Institutes (NSI). From a frequentist perspective, to the best of knowledge, there is no SAE literature on the estimation of Gini coefficients at the regional level using area-level data, and specifically with application of the FH model. The possible advantages of using a frequentist approach are, that it is probably easier to follow for common users who are more used to frequentist regression models and the available software for SAE methods implements mostly frequentist methods. In addition, there are a number of elaborated results from a frequentist perspective for the FH model that can be adapted. As the FH model allows for the use of a transformation, it is a common approach to satisfy the normality assumptions of the error terms or to ensure that the estimated values are within a predefined range. Slud and Maiti (2006), for example, propose a log-transformed FH model for skewed data, and in the case of proportions, for example Casas-Cordero et al. (2016) use an arcsine-transformed FH model to estimate poverty rates and Schmid et al. (2017) for literacy rates. To estimate Gini coefficients using the FH model, in this work the approach of Fabrizi and Trivisano (2016) is followed and a logit transformation is used to link

the response values to the related covariate information. This is also motivated by the condition that the estimated Gini coefficients must lie between zero and one, in addition to stabilizing the variance of the direct estimator and to promoting the normal distribution of the sampling errors and random effects of the model. The choice of a logit-normal rather than a beta likelihood as in Fabrizi and Trivisano (2016), is also driven by the possibility to use already existing results, such as those of Sugasawa and Kubokawa (2017) for the back-transformation. When using transformations the resulting point estimate is on the transformed scale and has to be back-transformed. An application of the inverse usually introduces a bias for nonlinear transformations, therefore Sugasawa and Kubokawa (2017) propose a bias-corrected back-transformation for general parametric transformations. This bias-corrected back-transformation is adopted to the logit transformation in this paper. Instead of the logit transformation, any other transformation could in principle also be used, as long as the inverse maps into a range between 0 and 1. For example, a complementary log-log or probit transformation could also be used if suitable transformations are available for the variance of the direct estimator and the back-transformation of the point estimator. In this paper, however, the focus is on the logit transformation, since it is one of the most common. To evaluate the accuracy of model-based SAE estimators, uncertainty measures must be estimated. As a common practice, the MSE is considered for this purpose. If analytical solutions for its estimation cannot be derived, bootstrap methods are often implemented instead. Here, the uncertainty of the estimated Gini coefficients is assessed using a bootstrap procedure following Gonzalez-Manteiga et al. (2008b) with an additional step of applying the bias-corrected back-transformation similar to Hadam et al. (2020). The validity of the presented point estimator using a logit-transformed FH model with a bias-corrected back-transformation, as well as that of the uncertainty measure, is demonstrated in a simulation study.

The paper is organized as follows. Section 5.2 describes the data used to illustrate the proposed methodology, in particular survey data from the Socio-Economic Panel (SOEP) and auxiliary data from administrative sources, such as the Census 2011 in Germany. The statistical methodology is introduced in Section 5.3. The validity of the proposed methodology is assessed in a simulation study in Section 5.4. Section 5.5 presents the application of the model-based small area method to estimate Gini coefficients for German regions. Section 5.6 completes the paper with some concluding remarks and discusses further potential research.

5.2 Sources of data and initial analysis

In this section, the data sources used for the analysis in Section 5.5 are described. Specifically, data from the German SOEP (Socio-Economic Panel, 2019) are used to form the target indicator, and data from the 2011 Census and the regional data base from the National Statistical Office are taken as auxiliary information. To have both data sources from the same year, the SOEP data collected in 2011 are used. Furthermore, a preliminary calculation of the Gini coefficients at a regional level is presented.

5.2.1 German Socio-Economic Panel

The German SOEP is a longitudinal study that has been running since 1984 and is conducted annually. It currently covers about 15,000 private households in Germany and aims to represent German society. Information is collected on various areas of life, such as demography, employment, taxes, income, education, health and satisfaction. The SOEP-team at the German Institute for Economic Research (DIW Berlin) prepares and provides the survey data. The main dataset SOEP Core currently consists of 12 sub samples. The initial sample, sample A, was first surveyed in 1984 and represents the West German population of the Federal Republic of Germany (Kara et al., 2019). In 1990, the initial sample East after the reunification was included, representative of the East German population of the German Democratic

Republic. Over the years (1998, 2000, 2002, 2011), four refreshment samples were added, further enlarging the total sample. In addition to the refreshment samples, other special samples to increase statistical power were included, such as the migration samples in 1984 and 1994/95, which oversamples foreigners or the high income sample in 2002 to represent households at the top of the income distribution. Sampled households are surveyed every year. The SOEP questionnaires are constructed in such a way that individuals in a SOEP household can be studied from birth to adulthood and over the rest of their lives. The SOEP aims to measure stability and identify changes across time, so the survey methodology remains almost identical over time (Kara et al., 2019). In the analysis in Section 5.5, data from the available refreshment sample in survey year 2011 is used. The sample aimed to cover a cross-section of private German households and is based on a clustered sampling strategy. Households were drawn at random from 307 primary sampling units (PSU) stratified by federal states, administrative regions and a classification of municipalities by number of inhabitants (Siegers et al., 2020). A random walk procedure was applied to select the addresses within each PSU. The provided household weights account for sampling design, non-response, and panel attrition and are further post-stratified to known population distributions based on the German microcensus.

The Gini coefficients calculated in this paper are computed with household-level data. The variable to form the target indicator in this section and for the application in Section 5.5 is the equivalised disposable household income, which is calculated using total net household income divided by equivalised household size. The equivalised household size is derived using the Organisation for Economic Co-operation and Development (OECD) modified scale first proposed by Hagenaars et al. (1994). The distribution of the variable in the sample is reported in Table 5.1. The Gini coefficient for the equivalised disposable household income reported in 2011 for Germany by OECD (2011) is 0.29. Goebel and Frick (2005) investigate regional income inequality by estimating Gini coefficients for East and West Germany and for a further regional stratification by dividing Germany into northern, eastern, western and southern states. This analysis indicates that there is regional heterogeneity in income inequality. In addition, the OECD reports Gini coefficients for the German federal states (OECD, 2013), which reveals further regional differences in inequality ranging from 0.23 in Saxony to 0.32 in Hesse. Another spatial dis-aggregation that enables the examination of inequality in rural and urban regions is the consideration of 96 spatial planning regions (SPRs) of the Federal Office for Building and Regional Planning. SPRs are composed of several administrative districts and form an intermediate regional level between these districts and the federal states. A map showing the assignment of the SPRs and associated labels can be found in the Appendix in Figure E.1 and Table E.1. The information to which SPR the residence of a SOEP household is assigned to can be found in the SOEP geocodes (Goebel, 2020). The investigation of regional differences in income inequality in Germany is therefore done for the 96 SPRs. Figure 5.1 shows estimated Gini coefficients from left to right for East and West Germany, a fourfold division of Germany into East, North, South and Central, the federal states and the SPRs. The first two maps already show that there are regional differences, as illustrated by Goebel and Frick (2005). The map of the federal states underlines this heterogeneity. Looking at the fourth map, the regional differences in income inequality become even more obvious. At the same time looking at Table 5.1, for some

Table 5.1: Distribution of equivalised disposable household income [€], sample sizes for SPRs and number of SPRs without observations.

	Min	1stQ	Median	Mean	3rdQ	Max	No obs.
Equal. disp. income	0	12363	17805	20579	25270	322508	
SPR sample size	4	17	27	35	47	153	7

SPRs, these estimates are based on a very small sample size, so that the reliability of the estimates cannot be guaranteed. To improve the accuracy of estimated Gini coefficients for SPRs with small sample

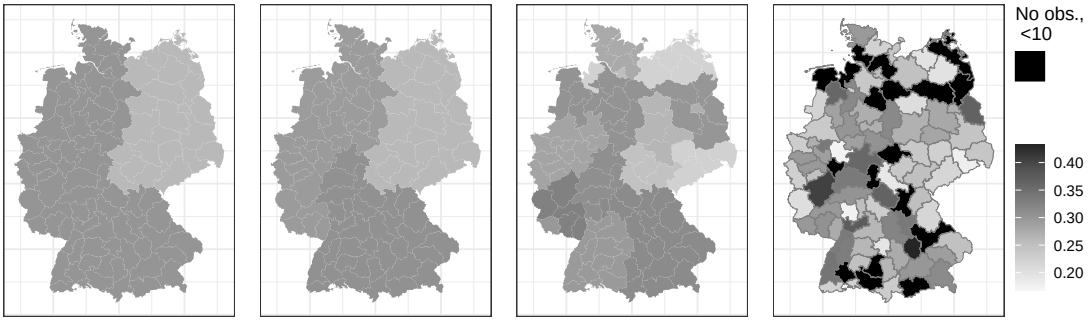


Figure 5.1: Gini coefficients for equivalised disposable income for East and West Germany (left), a fourfold division of Germany into East, North, South and Central, the federal states and SPRs (right). SPRs with no or less than 10 observations are colored in black.

sizes, model-based SAE methods combine direct estimates with auxiliary information from registers by statistical models. Furthermore, those methods allow to provide estimates for regions that have no observations in the survey, usually referred to as out-of-sample (OOS) regions. This is the case for 7 SPRs. According to the privacy agreement with the data provider, direct estimates of SPRs with less than 10 observations cannot be reported. This applies to 11 SPRs. In the map for the SPRs (Figure 5.1), these and the OOS SPRs are colored in black.

5.2.2 Auxiliary information

To improve the accuracy of the target indicator, the model described in Section 5.3.1 makes use of auxiliary information from administrative data sources as registers or census data at an aggregated level. For the application in this work, German Census data from 2011 (Statistische Ämter des Bundes und der Länder, 2011a) is used, which is publicly available at an administrative district level. Furthermore data on taxes, gross domestic product (GDP), mortality and birth numbers available from the National Statistical Offices are used (Statistische Ämter des Bundes und der Länder, 2011c). A detailed explanation of the calculation of the GDP on district level can be found in Statistische Ämter der Länder (2021). To obtain the data at the same level as the survey data, they are aggregated to SPR level. The assignment of counties and districts to SPRs is provided by the Federal Office for Building and Regional Planning (Bundesinstitut für Bau-, Stadt-, und Raumforschung, 2017). The objective is to find variables in the data that are related to income inequality and could serve as possible predictors. Furceri and Ostry (2019) examine robust drivers of income inequality and identify, among other factors, the level of development and demographics as key determinants, as well as the extent of unemployment. Perugini and Martino (2008) examine the factors that drive inequality within European regions. Both divide the factors into groups of demographic, institutional and economic condition variables, among others. The possible covariates that were able to be extracted and aggregated from the data sources available are presented in Table 5.2 with summary statistics. Although Furceri and Ostry (2019) consider inequality determinants between countries, this could be transferred to within country inequality and development. When considering economic conditions, in addition to GDP, which is a measure of a region's development, the shares of the agricultural, industrial and social service sectors in the labor market are also an indicator of economic development. Since the industrial sector is generally expected to generate higher income, this could lead to a better distribution of income than a high share in the agricultural sector. In line with Fabrizi and Trivisano (2016) and Perugini and Martino (2008) taxable income and the share of income taxpayers can be an indirect measure of labor performance and, moreover, an indicator of the resources that local governments could use to fund education, child care, health, etc., to foster fu-

Table 5.2: Distributions of possible auxiliary information.

	Min	1stQ	Median	Mean	3rdQ	Max
Economic/ Institutional conditions						
GDP per resident [€]	22159	72625	121001	132171	174988	381263
log(GDP per resident)	10.010	11.190	11.700	11.630	12.070	12.850
Avg. taxable income per person in Tsd. [€]	2.972	4.192	4.933	4.886	5.553	8.840
Share income tax payer	0.399	0.457	0.481	0.480	0.499	0.683
Share agricultural employment sector	0.000	0.002	0.008	0.010	0.015	0.039
Share industrial employment sector	0.100	0.181	0.216	0.219	0.253	0.359
Share service sector	0.452	0.512	0.542	0.546	0.572	0.664
Unemployment ratio	0.002	0.021	0.035	0.038	0.053	0.087
High education ratio	0.153	0.247	0.293	0.297	0.340	0.488
Demographics						
Population density	44.0	117.5	178.0	330.7	274.5	3927.0
log(Population density)	3.784	4.766	5.182	5.347	5.615	8.276
Foreign residents ratio	0.009	0.033	0.054	0.060	0.085	0.153
Child dependency ratio	0.162	0.194	0.206	0.204	0.220	0.241
Elderly dependency ratio	0.263	0.291	0.316	0.318	0.338	0.415
Births rate	6.734	7.455	7.814	7.922	8.280	11.837
Mortality rate	8.000	9.701	10.691	10.778	11.677	14.358

ture growth and thus reduce inequality. The level of unemployment naturally measures the economic situation of a region, just as the level of education is a proxy for development. An approach similar to Fabrizi and Trivisano (2016) is used to calculate a high education ratio. Therefore the number of people aged between 18 and 64 with at least high school diploma are divided by the number of all people aged between 18 and 64. Following Furceri and Ostry (2019), demographic data such as dependency ratios, birth, and death rates are also among the possible covariates, as they indirectly approximate economic development. This is also true for the foreigner rate, as immigration could lead to an increasing wage gap (Furceri and Ostry, 2019).

5.3 Small area estimation method

In this section, the statistical methodology is presented. The underlying model for estimating small area means was proposed by Fay and Herriot (1979), which combines aggregate population auxiliary variables with direct estimators based on survey data. In this work, the target indicators are area-specific Gini coefficients. Since it is a nonlinear indicator within a specified range, a logit transformation is applied to promote the normality assumption of the model and to ensure that the estimates are between 0 and 1. To measure the uncertainty of the point estimator, a parametric bootstrap procedure is presented.

5.3.1 Logit-transformed Fay-Herriot model

Let N be the size of a finite population divided into $d = 1, ..., D$ domains and n the sample size with $i = 1, ..., n_d$ units per domain so that $n = \sum_{d=1}^{D} n_d$. The FH model is a two-level model that includes a sampling model at the first level, assuming that the direct estimator consists of the true domain-specific population indicator θ_d and sampling errors e_d:

$$\hat{\theta}_d^{Dir} = \theta_d + e_d, \quad e_d \overset{ind}{\sim} \mathcal{N}\left(0, \sigma_{e_d}^2\right). \tag{5.1}$$

The sampling errors e_d are assumed to be independently normally distributed with known variance $\sigma_{e_d}^2$. However, although the sample variance $\sigma_{e_d}^2$ is taken as known, in many applications it has to be estimated itself, what can be done on the basis of unit-level sample data (Rivest and Vandal, 2002;

Wang and Fuller, 2003; You and Chapman, 2006) or by bootstrap algorithms proposed in Alfons and Templ (2013). There are several proposed direct estimators for the Gini coefficient in the literature. A common estimator is the one proposed by Alfons and Templ (2013). Fabrizi and Trivisano (2016) show in a simulation experiment, that this estimator can have a negative bias when sample sizes are small and propose a corrected version with a bias reduction. The direct estimator proposed by Fabrizi and Trivisano (2016) is defined as

$$\hat{\theta}_d^{Dir} = \frac{1}{2\hat{\bar{Y}}_d} \frac{\sum_{i=1}^{n_d} \sum_{j=1}^{n_d} w_{di} w_{dj} |y_{di} - y_{dj}|}{\hat{N}_d^2 - \sum_{i=1}^{n_d} w_{di}^2}, \tag{5.2}$$

with $\hat{N}_d = \sum_{i=1}^{n_d} w_{di}$ and $\hat{\bar{Y}}_d = \hat{N}_d^{-1} \sum_{i=1}^{n_d} w_{di} y_{di}$, where y_{di} is the income or wealth variale, in this paper the equivalised disposable household income and w_{di} denote the sampling weights. By including the sample weights in the associated variance estimate, the direct estimator incorporates the complex design information. The variances $\sigma_{e_d}^2$ of $\hat{\theta}_d^{Dir}$ for $d = 1, ..., D$ can be estimated via a naive or calibrated bootstrap procedure described in Alfons and Templ (2013). Since the direct variance estimates are based on small sample sizes a variance smoothing model analogous to that in Fabrizi and Trivisano (2016) is used for stabilization. The model assumes a beta distribution for the Gini coefficient and uses the relationship between the expected value and the variance of the beta distribution. It is defined as follows:

$$\frac{\hat{\theta}_d^{Dir^2}(1 - \hat{\theta}_d^{Dir^2})}{2\sigma_{e_d}^2} = \lambda n_d + \epsilon_d \tag{5.3}$$

where the error term is assumed to be normally distributed $\epsilon_d \sim \mathcal{N}(0, \tau^2)$ and λ is estimated using least squares.

The second level of the FH model is a linking model that links covariate information to the population indicator. x_d is a $p \times 1$ vector of domain-specific population covariates and β is the corresponding $p \times 1$ vector of regression coefficients. v_d are domain-specific random effects, which are normally distributed:

$$\theta_d = x_d^T \beta + v_d, \quad v_d \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_v^2\right). \tag{5.4}$$

To ensure that the estimated Gini coefficients lie within $(0, 1)$, to further stabilize the variance and following Fabrizi and Trivisano (2016), the logit function is applied to the direct estimator from Equation (5.2):

$$\hat{\theta}_d^{Dir*} = \text{logit}(\hat{\theta}_d^{Dir}) = log\left(\frac{\hat{\theta}_d^{Dir}}{(1 - \hat{\theta}_d^{Dir})}\right).$$

In the following, * always refers to the logit-scale. To obtain the variances of the direct estimator on the transformed scale, one can transfer the smoothed bootstrap variances to the logit scale using Taylor expansion for moments, which leads to:

$$\sigma_{e_d}^{2*} = \frac{\sigma_{e_d}^2}{\left[\hat{\theta}_d^{Dir}\left(1 - \hat{\theta}_d^{Dir}\right)\right]^2}. \tag{5.5}$$

Using a Taylor expansion for moments to transform variances from the original scale to the transformed scale is a common procedure in SAE as in Neves et al. (2013) and Council (2000).

The combination of the sampling model in (5.1) and the linking model in (5.4) with the logit-transformed direct estimator results in:

$$\text{logit}\left(\hat{\theta}_d^{Dir}\right) = x_d^T \beta + v_d + e_d^*, \quad v_d \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_v^2\right), \quad e_d^* \overset{ind}{\sim} \mathcal{N}\left(0, \sigma_{e_d}^{2*}\right). \tag{5.6}$$

The unknown parameters of the model (5.4) to be estimated are the model variance σ_v^2 and the regression

coefficients β. Methods to estimate σ_v^2 are for example restricted maximum likelihood (REML), maximum likelihood (ML) and the FH method-of-moments. Details on model variance estimation methods can be found, for example, in Rao and Molina (2015). In this paper, the REML method is used, which has the advantage over the ML method of taking into account the loss of degrees of freedom in the estimation of the regression coefficients β (Rao and Molina, 2015). Let $\hat{\sigma}_v^2$ be an unbiased estimator for σ_v^2. Then the best linear unbiased estimator (BLUE) under model (5.6) for the regression coefficients β is given by:

$$\hat{\beta} = \hat{\beta}\left(\hat{\sigma}_v^2\right) = \left(\sum_{d=1}^{D} \frac{x_d x_d^T}{\sigma_{e_d}^{2*} + \hat{\sigma}_v^2}\right)^{-1} \left(\sum_{d=1}^{D} \frac{x_d \hat{\theta}_d^{Dir*}}{\sigma_{e_d}^{2*} + \hat{\sigma}_v^2}\right).$$

Since the model inputs are on the logit scale, the estimated regression coefficients $\hat{\beta}$ as well. Therefore, only the direction of the effect on the estimated model-based Gini coefficient on the original scale can be interpreted.

The FH estimator on the logit scale is obtained by:

$$\hat{\theta}_d^{FH*} = x_d^T\hat{\beta} + \hat{v}_d = \hat{\gamma}_d\hat{\theta}_d^{Dir*} + (1-\hat{\gamma}_d)\, x_d^T\hat{\beta} \quad \text{with} \quad \hat{\gamma}_d = \frac{\hat{\sigma}_v^2}{\sigma_{e_d}^{2*} + \hat{\sigma}_v^2}. \tag{5.7}$$

$\hat{\gamma}_d$ is the shrinkage factor which determines an optimal balance between the direct estimator and the synthetic component. If the variance of the direct estimator is large, more weight is given to the synthetic component. The estimated model variance, i.e., the variance of the random effects σ_v^2, is also on the logit scale, as are the sampling variances. Therefore, the weighting factor can also be interpreted as the proportion of the variation explained by the hierarchical structure of the data. For highly skewed data, the transformation helps to better fit the linear relationship in the model, so using a transformation on skewed data can often give more weight to the synthetic part. Since the direct estimators and their variances of the Gini coefficients were transformed to the logit scale as model input for the FH model, the resulting FH estimator $\hat{\theta}_d^{FH*}$ of the Gini coefficients is also still on the logit scale. To obtain the estimates on the original scale, a back transformation is required. As naive inverse back-transformations (in this case the logistic function) usually introduce a bias for nonlinear functions, Sugasawa and Kubokawa (2017) present an asymptotically unbiased back-transformation for a general parametric transformation. Hadam et al. (2020) applies this to the arcsine transformation, for example. Following Sugasawa and Kubokawa (2017) to obtain a bias-corrected back-transformation for $\hat{\theta}_d^{FH}$, the normal distribution of the transformed FH estimator on the logit-scale and the expected value (E) of a transformation (here the inverse logit) are used. The bias-corrected back-transformation applied to obtain the final FH estimates of the Gini coefficients $\hat{\theta}_d^{FH}$ at the original scale is as follows:

$$\begin{aligned}\hat{\theta}_d^{FH} &= \mathrm{E}\left[\text{logit}^{-1}\left(\hat{\theta}_d^{FH*}\right)\right] = \mathrm{E}\left[\frac{exp\left(\hat{\theta}_d^{FH*}\right)}{1+exp\left(\hat{\theta}_d^{FH*}\right)}\right] = \int_{-\infty}^{\infty} \frac{exp\,(t)}{1+exp\,(t)} f_{\hat{\theta}_d^{FH*}}(t)\, dt \\ &= \int_{-\infty}^{\infty} \frac{exp\,(t)}{1+exp\,(t)} \frac{1}{\sqrt{2\pi\frac{\hat{\sigma}_v^2\sigma_{e_d}^{2*}}{\hat{\sigma}_v^2+\sigma_{e_d}^{2*}}}} exp\left(-\frac{\left(t-\hat{\theta}_d^{FH*}\right)^2}{2\frac{\hat{\sigma}_v^2\sigma_{e_d}^{2*}}{\hat{\sigma}_v^2+\sigma_{e_d}^{2*}}}\right) dt.\end{aligned} \tag{5.8}$$

In Equation (5.8) the integral has to be solved by numerical integration methods. The advantage of the bias-corrected back-transformation over the naive inverse is illustrated in the simulation experiment in Section 5.4.

5.3.2 Uncertainy measure

In order to evaluate the accuracy of the FH estimator with a logit transformation and to demonstrate the benefit of model-based estimators over direct ones, it is necessary to determine the degree of uncertainty. In the case of the FH estimator without a transformation, analytical solutions exist to estimate the MSE, such as the MSE estimator according to Prasad and Rao (1990). In the log-transformed FH model, Slud and Maiti (2006) also derived an analytical MSE estimator. There, the relationships between the log-normal distribution and the normal distribution and their expected values are used. This approach cannot be straightforwardly applied to the logit transformation and the relationship between the logit-normal and the normal distribution, as there are no analytical solutions for the moments of the former. A common approach to estimating the MSE if no analytical estimator can be derived is to use a bootstrap algorithm. In line with Gonzalez-Manteiga et al. (2008b), the MSE of $\hat{\theta}_d^{FH}$ is approximated with the following parametric bootstrap procedure:

1. Estimate the regression synthetic components $\hat{\beta}$ and $\hat{\sigma}_v^2$ using the direct components $\hat{\theta}_d^{Dir*}$ and $\sigma_{e_d}^{2^*}$ on the logit-scale.
2. For $b = 1, ..., B$
   (a) Generate sampling errors $e_d^{*(b)} \overset{ind}{\sim} \mathcal{N}\left(0, \sigma_{e_d}^{2^*}\right)$ and random effects $v_d^{(b)} \overset{iid}{\sim} \mathcal{N}\left(0, \hat{\sigma}_v^2\right)$.
   (b) Simulate a bootstrap sample $\hat{\theta}_d^{Dir*(b)} = x_d^T\hat{\beta} + v_d^{(b)} + e_d^{*(b)}$.
   (c) Calculate the true bootstrap population parameter $\theta_d^{*(b)} = x_d^T\hat{\beta} + v_d^{(b)}$ on the transformed scale and back-transform with $\theta_d^{(b)} = \frac{exp\left(\theta_d^{(b)}\right)}{1+exp\left(\theta_d^{(b)}\right)}$.
   (d) Estimate the bootstrap estimator of the model variance $\hat{\sigma}_v^{2(b)}$ using $\hat{\theta}_d^{Dir*(b)}$ and $\sigma_{e_d}^{2^*}$.
   (e) Using $\hat{\sigma}_v^{2(b)}$ and $\hat{\theta}_d^{Dir*(b)}$, estimate bootstrap estimators of the regression coefficients $\hat{\beta}^{(b)}$ and update the random effects $v_d^{(b)}$.
   (f) Determine the bootstrap estimator $\hat{\theta}_d^{FH*(b)}$ with Equation (5.7) by using the estimates from the previous step and back-transform to the original scale by applying the bias-corrected back-transformation from Equation (5.8) to obtain $\hat{\theta}_d^{FH(b)}$.
3. Estimate the MSE:

$$\widehat{\text{MSE}}(\hat{\theta}_d^{FH}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_d^{FH(b)} - \theta_d^{(b)}\right)^2. \tag{5.9}$$

The performance of the presented bootsrap MSE estimator is evaluated in the simulation experiment in Section 5.4.

5.3.3 An alternative estimator from a Bayesian perspective

As an alternative to the proposed methodology from a frequentist perspective Fabrizi and Trivisano (2016) presented a Bayesian Beta-regression model to get model-based estimators for the Gini concentration coefficients for small regions. This estimator is used in the simulation experiment in Section 5.4 as a comparative estimator. For a better understanding it is shortly introduced in the following. The sampling model with a Beta distribution as the underlying distribution for the direct estimator from Equation (5.2) for $d = 1, ..., D$ domains is defined as follows:

$$\hat{\theta}_d^{Dir} \sim Beta\left(\frac{2\phi_d}{1+\theta_d} - \theta_d, \frac{2\phi_d - \theta_d(1+\theta_d)}{1+\theta_d}\frac{1-\theta_d}{\theta_d}\right),$$

with expected value $E(\hat{\theta}_d^{Dir}|\theta_d) = \theta_d$ and variance $V(\hat{\theta}_d^{Dir}|\theta_d) = 2\hat{\phi}_d^{-1}\theta_d^2(1+\theta_d^2)$, where ϕ_d is the precision parameter of the Beta distribution and can be estimated from the survey data and the variances

of the direct estimator $\hat{\theta}_d^{Dir}$, which are assumed to be known here as well, inline with SAE literature. Using the variance smoothing model from Equation (5.3) ϕ_d can be estimated by $\hat{\phi}_d = \hat{\lambda} n_d$. For further details it is referred to Fabrizi and Trivisano (2016). The linking model with a logit link is defined as follows:

$$\text{logit}(\theta_d) = x_d^T \beta + v_d, \tag{5.10}$$

where x_d is a $p \times 1$ vector of domain-specific population covariates, β the corresponding $p \times 1$ vector of regression coefficients and v_d are the domain-specific random effects. To estimate the model in Equation (5.10) the specification of prior distributions for the random effects v_d, their variance σ_v^2 and the regression coefficients β are necessary. For β a normal prior with zero mean and large variances can be suggested: $\beta \sim N(0, kI)$, with $k = 100$ and I is the $p \times p$ identity matrix. For the random effects and their variance various prior specifications are possible. In the simulation experiment in Section 5.4 the following prior distribution is assumed because it proved to be preferable to other prior distributions according to Fabrizi and Trivisano (2016): $v_d \sim N(0, \sigma_v^2)$ with $\sigma_v^2 \sim \text{half-}t(\nu = 3, \text{A} = 1)$, where ν are the degrees of freedom and A is the scale parameter. For the other possible specifications it is referred to Fabrizi and Trivisano (2016). The posterior distributions of the Gini coefficients are approximated by a MCMC algorithm, from which one directly obtains the point estimate for θ_d and a corresponding uncertainty measure, usually the expected value and variance of the posterior distribution given the data.

5.4 Simulation study

To evaluate the performance of the proposed estimators in Section 5.3 in terms of bias and accuracy, a model-based simulation experiment is conducted. In particular, the performance of the point estimator compared to three alternative estimators is of interest, as well as the presented uncertainty measure. The simulation setup is based on the estimated parameters from Section 5.5 and was chosen to mimic real data. The data are created for $D = 89$ domains. For the data generation process of the true parameter of interest and its direct estimator, the model variance and sampling variances from the SOEP data from Section 5.5 are used. The true parameters of interest θ_d for $d = 1, ..., 89$ domains are derived via $\text{logit}(\theta_d) = \beta_0 + \beta_1 x + v_d$ with $\beta_0 = -1.5$, $\beta_1 = 1$ and covariate $x \sim \mathcal{LN}(-0.5, 0.04)$ generated so that the true values lie in a range of realistic Gini coefficients. The random effects v_d follow a normal distribution $\mathcal{N}(0, 0.029)$, where the variance parameter equals the estimated model variance in Section 5.5.1. The direct estimates are generated as $\text{logit}(\hat{\theta}_d^{Dir}) = \beta_0 + \beta_1 x + v_d + e_d$, with $e_d \sim \mathcal{N}(0, \sigma_{e_d}^2)$ where $\sigma_{e_d}^2$ are the direct variances on the logit-scale of the 89 observed SPRs from Section 5.5. They are listed in Table E.2 in the Appendix. The distributions of the given and resulting parameters in the simulation are reported in Table 5.3. The data scenario was generated for $R = 1,000$ simulation runs.

Table 5.3: Summary of parameters in the simulation setting.

	Min	1stQ	Median	Mean	3rdQ	Max
θ_d	0.206	0.264	0.298	0.294	0.319	0.396
$\hat{\theta}_d^{Dir}$	0.141	0.250	0.295	0.293	0.332	0.448
$\sigma_{e_d}^2$	0.082	0.149	0.188	0.206	0.230	0.589
γ_d	0.077	0.356	0.454	0.453	0.567	0.813
x	0.547	0.583	0.598	0.601	0.616	0.670

The performance of the proposed bias-corrected estimator from Equation (5.8), denoted by logit FH.bc, is evaluated in comparison to three estimators: To a logit-transformed FH estimator with a naive back-transformation using the inverse of the logit function (logit FH.naive), to the usual FH estimator (FH), and to the estimator proposed by Fabrizi and Trivisano (2016) and shortly introduced in Section 5.3.3. In the MCMC algorithm for the latter, a sample of 10,000 draws, with a preceding burn-in phase of

20,000 draws was used and the code provided by Fabrizi et al. (2016) was utilized to implement the estimator. The performance of the estimators is assessed by the distribution over the domains of the domain-specific absolute bias (ABias) and root mean squared error (RMSE), given as follows:

$$\text{ABias}(\hat{\theta}_d) = \left| \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}_{d_r} - \theta_{d_r} \right) \right|, \qquad \text{RMSE}(\hat{\theta}_d) = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}_{d_r} - \theta_{d_r} \right)^2}, \tag{5.11}$$

where $\hat{\theta}_{d_r}$ denotes the estimator of the target indicator in domain d and replication r and θ_{d_r} the true value. Table 5.4 reports the distributions of the domain-specific ABias and RMSE over domains for

Table 5.4: Summary over domains of absolute bias and RMSE.

	Estimator	Min	1stQ	Median	Mean	3rdQ	Max
$10^3\times$ ABias	Bayesian Beta	0.005	0.241	0.528	0.690	1.126	2.398
	FH	0.099	0.686	1.408	1.457	2.033	4.024
	logit FH.bc	0.031	0.238	0.548	0.589	0.840	1.790
	logit FH.naive	0.042	0.338	0.778	0.827	1.188	2.447
$10^3\times$ RMSE	Bayesian Beta	15.320	23.260	26.290	26.330	28.970	34.680
	FH	15.220	23.200	26.940	26.630	29.320	36.730
	logit FH.bc	15.270	23.150	26.240	26.210	28.830	34.250
	logit FH.naive	15.260	23.160	26.260	26.220	28.830	34.270

the evaluated estimators. Starting with the bias it can be noted that the estimators, which use a logit transformation (Bayesian Beta, logit FH.bc and logit FH.naive) outperform the FH estimator (FH) without a transformation, which is a natural result due to the data generating process. Looking specifically at logit FH.bc and logit FH.naive, the reduction in bias due to the bias-corrected back-transformation is noticeable across the entire range of the distribution. Comparing the two median values, the use of logit FH.bc resulted in a 30% reduction in the median value of logit FH.naive. Further the results of the proposed bias-corrected estimator are comparable to those of the Bayesian estimator. In terms of efficiency, the four estimators provide very similar results with negligible differences. It is worth mentioning here that the bias-corrected back-transformation does not lead to a loss of efficiency and that the performance is similar to that of the Bayesian estimator proposed by Fabrizi and Trivisano (2016). Since in the data generating process the logit transformation is used, the comparison of the three estimators which use a logit-link is in that sense fair, that this refers to their use-case. Furthermore, the simulated direct estimators lie within a range of realistic values for the Gini coefficients, and are not at the edges of the distribution, where a higher gain of the bias-corrected back-transformation compared to the naive can be expected. Only the comparison to the standard FH estimator is somewhat unfair, since the data scenario does not fit the untransformed FH model. Nevertheless, the comparison is of interest, since this approach corresponds to the simplest and is mainly used in practice. To investigate whether the differences between the methods are a result of the SAE estimators themselves or may be within a simulation-induced margin of error, the Monte Carlo error (MCE) is estimated with a Jackknife estimator following Koehler et al. (2009). The distributions of MCEs of the quantities of interest presented in Equation (5.11) are given in Table E.3 in the Appendix. Since the distributions across the domains of each method per quantity are very similar, it can be concluded that the differences from Table 5.4 are effective and not attributable to a MCE.

Next, the bootstrap MSE estimator from Equation (5.9) is examined for the estimator defined in Equation (5.8). It is denoted by $\widehat{\text{MSE}}_{d_r}$ for domain d of simulation run r. The estimator was calculated with $B = 500$ bootstrap replications in each simulation run. Its performance is evaluated comparing the estimated and the RMSE defined in Equation (5.11), which is treated as the true RMSE. As a measure

of bias the relative bias (RB RMSE) is chosen, which is defined as follows:

$$\text{RB RMSE}(\hat{\theta}_d) = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\widehat{\text{MSE}}_{d_r}} - \text{RMSE}(\hat{\theta}_d)}{\text{RMSE}(\hat{\theta}_d)}.$$

Table 5.5 reports the distributions of the domain-specific RB RMSE over domains. It can already be seen that the percentage values are within an acceptable and common range for MSE estimators with a median relative bias of -1.1%. To have a closer look on the performance of the bootstrap MSE estimator

Table 5.5: Summary over domains of relative bias of estimated RMSE of logit FH.bc.

	Min	1stQ	Median	Mean	3rdQ	Max
RB RMSE [%]	-8.746	-3.434	-1.132	-0.836	1.450	8.283

with a bias-corrected back-transformation the estimated and true RMSE values per domain are plotted in Figure 5.2. The domains are ordered by decreasing sampling variances, which were used to construct the direct estimators. First, it can be observed that as the sampling variance decreases, the true RMSE also decreases, since a lower sampling variance is usually associated with a higher sample size and thus a lower RMSE. Second, the estimated RMSE tracks this behavior very well and thus captures the true uncertainty of the estimate in this setting. In summary, the bias-correction in the back-transformation is advantageous over the naive back-transformation in the given setting based on real data. Furthermore, the bootstrap MSE estimator leads to good results and provides a good estimate for the uncertainty.

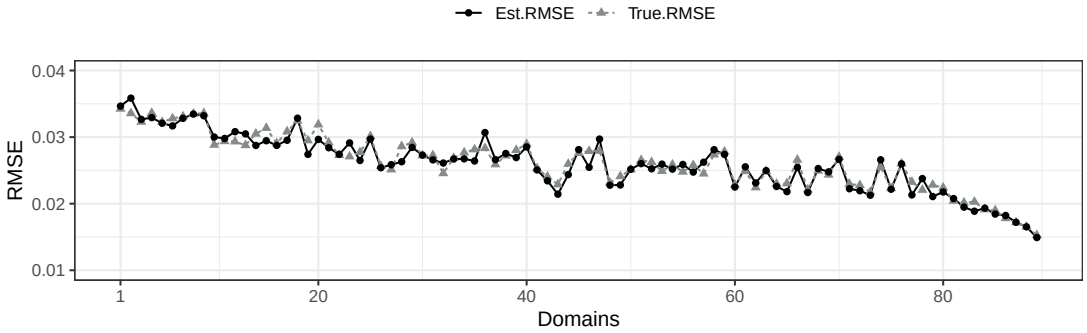


Figure 5.2: Estimated and true RMSE of logit FH.bc. Domains are ordered by decreasing sampling variances.

5.5 Application to German spatial planning regions

In this section, the methodology presented in Section 5.3 is illustrated using the data described in Section 5.2. In particular, the logit-transformed FH model with a bias-corrected back-transformation is used to estimate Gini coefficients for German SPRs, which are the domains in this application. At the same time, the advantage of using model-based small area methods in terms of increased accuracy is demonstrated. The SOEP sample used here contains data for 89 out of 96 SPRs in Germany with a total sample size of about 3,100 households. In this application the Gini coefficients for the equivalised disposable household income are estimated. Since income distributions often have a heavy right-hand tail, the sensitivity of inequality measures to outliers based on those variables is discussed in Alfons et al. (2013) and Cowell and Flachaire (2007). The Gini coefficient is especially affected by extreme outliers and Alfons et al. (2013) therefore propose a Pareto tail modeling, which is also applied here. In this case, observations in the income distribution that are above a threshold, i.e. the scale parameter of the Pareto

distribution determined according to Van Kerm (2007), and are additionally extreme for the Pareto distribution are identified as outliers. These outliers are replaced by values of the underlying theoretical Pareto distribution. This approach was implemented by Alfons and Templ (2013) in the **laeken** R-package. In the whole sample, 65 households lie in the upper tail of the distribution of which in total two households from one SPR each (Cologne, Southern Upper Rhine) are identified as outliers and are replaced. The Gini coefficients for the SPRs are estimated using the direct estimator $\hat{\theta}_d^{Dir}$ from Equation (5.2) proposed by Fabrizi and Trivisano (2016). The sampling variances $\sigma_{e_d}^2$ are estimated with the naive bootstrap procedure according to Alfons and Templ (2013) and implemented in the R-package **laeken**. Following Fabrizi and Trivisano (2016) the variance smoothing model from Equation (5.3) was estimated to further smooth and stabilize the variances. Afterwards the smoothed sampling variances are brought to the logit scale with Equation (5.5).

5.5.1 Model selection and validation

Before moving to the discussion of model-based estimates of Gini coefficients obtained with Equations (5.7) and (5.8), the variable selection and testing of model assumptions using diagnostics is reviewed. From the set of possible covariates for predicting Gini coefficients and improving accuracy given in Table 5.2, reasonable covariates are selected using an approach developed especially for FH models. Marhuenda et al. (2014) discuss various methods for FH model selection which are variants of common criteria like the Akaike Information criterion (AIC) and Kullback symmetric divergence criterion (KIC) and argue that common AIC over-parameterize FH models. They conclude, that a KIC bootstrap variant (KICb2) is the best selection criterion for FH models. Therefore a step-wise selection procedure with KICb2 criterion proposed by Marhuenda et al. (2014) with $B = 300$ bootstrap replications was applied, which is implemented in the R-package **emdi** (Kreutzmann et al., 2019). The model selection was done with logit $\left(\hat{\theta}_d^{Dir}\right)$ as dependent variable and the transformed direct variances $\sigma_{e_d}^{2^*}$. The final model includes only the variable log(GDP per resident), which has an estimated positive effect. This is consistent with the hypothesis of Perugini and Martino (2008) that an increase in the regional level of development, with GDP serving as a proxy for economic development, promotes income inequality. The predictive power of the model is evaluated using an adjusted R^2 specifically for FH models proposed by Lahiri and Suntornchost (2015), which incorporates the variability of the sampling error. The model yields only a value of 16%, which is comparatively low, nevertheless the main goal of model-based small area methods, namely the gain in accuracy for small sample sizes, can be achieved, as can be seen in the next section. The model assumptions of normally distributed residuals and random effects are tested with the Shapiro-Wilk test and yield p-values of 0.854 and 0.147, respectively, thus normality cannot be rejected at a significance level of 5%. The model variance estimated using the REML method is $\hat{\sigma}_v^2 = 0.029$ and is used in Section 5.4 as part of the data generating process.

5.5.2 Gain in accuracy

Before looking at the model-based estimates of the Gini coefficients the gain in accuracy compared to the direct estimator is examined. The coefficients of variation (CV) per SPR for the proposed model-based estimator (logit FH.bc) and the direct estimator (Direct) are reported in Figure 5.3, where the SPRs are ordered by increasing sample sizes, starting with the OOS SPRs. The uncertainty of the bias-corrected logit-transformed FH estimator from Equation (5.8) is measured using the bootstrap algorithm presented in Section 5.3.2 with $B = 500$ bootstrap replications. The gain in efficiency is achieved for all SPRs as the CVs of the model-based estimators are always smaller than of the direct ones with a decreasing difference with higher sample sizes. This behavior is to be expected, as direct estimates become more reliable with higher sample sizes thus more weight is put on the direct component. For 13 of the 89 observed SPRs, the CV can be moved from above 20% to below this threshold using the

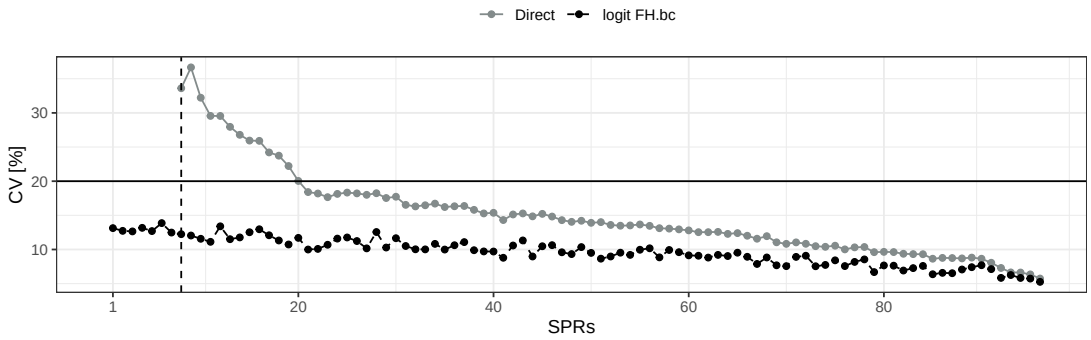


Figure 5.3: CVs of Direct and logit FH.bc. SPRs are ordered by increasing sample sizes, OOS SPRs first.

model-based estimator. The threshold of 20% is a common value up to which estimates are considered reliable (Eurostat, 2023). Table 5.6 shows the distribution of the estimated Gini coefficients and the

Table 5.6: Summary of point estimators and corresponding CVs [%] over SPRs, OOS SPRs in separate lines.

	Min	1stQ	Median	Mean	3rdQ	Max
Direct	0.1674	0.2313	0.2631	0.2706	0.3031	0.4321
logit FH.bc	0.2112	0.2484	0.2657	0.2691	0.2884	0.3568
logit FH.bc OOS	0.2428	0.2493	0.2503	0.2543	0.2614	0.2656
CV Direct	5.75	10.48	13.66	14.99	17.54	36.66
CV logit FH.bc	5.26	7.74	9.53	9.38	10.63	13.39
CV logit FH.bc OOS	12.48	12.68	12.73	12.96	13.15	13.86

corresponding CVs. The first observation is that the distribution of the direct estimator across SPRs is wider than that of the model-based estimator, while the mean and median values of the distribution correspond to each other. This is in line with the expectation that the model-based estimates should be consistent with the direct estimates but more precise. The expected shrinkage to the mean effect can additionally be seen in Figure 5.4, where the direct estimates are plotted against the model-based estimates. It can be observed that the SPRs with a low direct estimate correspond to a higher model-based estimate and vice versa, indicating the regression to the mean. Examination of the OOS SPRS in Table 5.6 shows that the point estimates lie in the middle of the distribution of model-based estimates for observed SPRs. The CVs are instead at the high end of the distribution, which makes sense considering that these observations were not used to estimate the model. To further investigate the quality of the model-based estimator, a closer look can be taken at Figure 5.5. There, the shrinkage factor $\hat{\gamma}_d$ from Equation (5.7), which indicates how much the direct component is weighted, is presented for each SPR with the corresponding sample size. On the x-axis are the SPRs ordered by decreasing sample sizes. It can be observed that in SPRs with higher sample sizes, the direct component is weighted more heavily, so that direct estimates and model-based estimates are very similar for SPRs with larger sample sizes. While the model-based estimator is more synthetic at smaller sample sizes.

5.5.3 Small area estimates

The regional distribution of the Gini coefficients estimated using the presented methodology for the 96 SPRs is mapped in Figure 5.6. The regional heterogeneity of income inequality within a region can be observed similar to the map in Figure 5.1. Figure 5.6 shows a similar pattern to Goebel and Frick (2005)

in that income inequality is still lower in eastern Germany than in the west, although different levels of inequality are estimated within the eastern regions. In the rural SPRs of the Northeast, inequality is lower than in the Baltic region. The estimated Gini coefficient of the SPR east of Berlin (Oderland-Spree) is relatively high compared to neighboring SPRs. This maybe due to a mixture of rural and urban SPRs next to Berlin and, according to Perugini and Martino (2008), to the coexistence of specific and mobile labor segments. Furthermore, taking into account the results of Immel and Peichl (2020) that in these regions the share of the lowest-income 40% of households is relatively high compared to the rest of Germany. Likewise, the share of the highest-income top 10% is not exceptionally low, probably due to proximity to Berlin. This mix could lead to higher income inequality. The estimated Gini coefficient for Berlin is 0.26, which is similar to the value of 0.28 reported by OECD (2013) for 2013. A more general result, that the northern regions of West Germany tend to have lower Gini coefficients than the regions in the south and center, could be explained by Immel and Peichl (2020)'s findings that disproportionately few of the top 10% income earners live in the north of West Germany and disproportionately many in the south. The highest estimated Gini coefficient is for the SPR Central Rhine-Westerwald, with the city of Koblenz at its center, surrounded by more suburban SPRs. According to Immel and Peichl (2020), the city of Koblenz has a relatively high share of top 10% highest income households, which could be the driver of income inequality in this region. In general, it can be noted that where Immel and Peichl (2020) identify a high share of the highest-income 10%, income inequality also tends to be rather high.

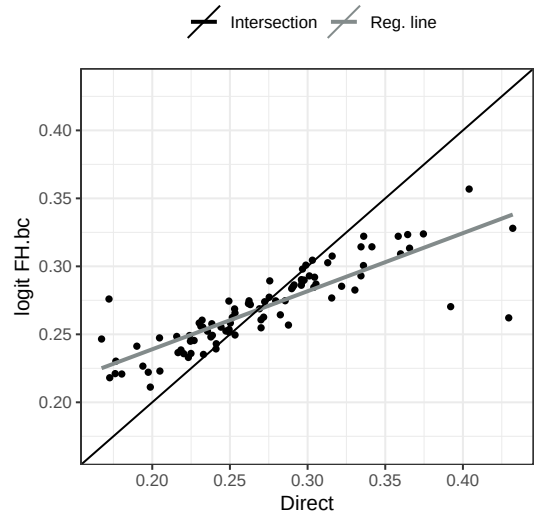


Figure 5.4: Direct vs. model-based estimated Gini coeffients.

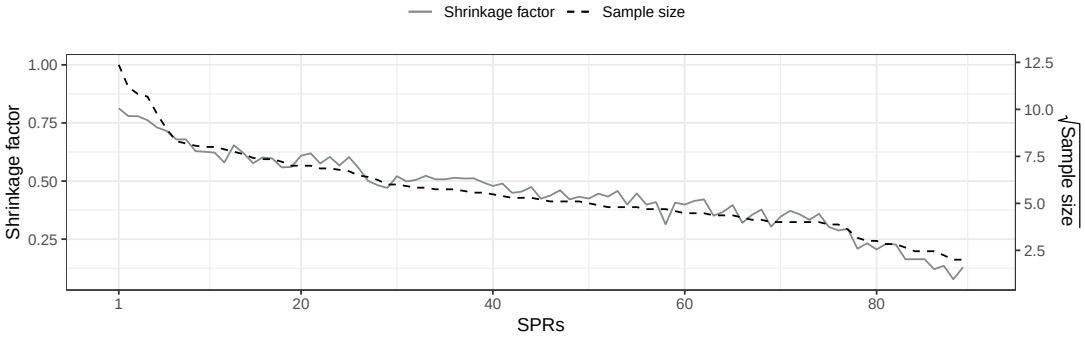


Figure 5.5: Shrinkage factor $\hat{\gamma}_d$ and sample sizes per SPR. SPRs are ordered by decreasing sample sizes.

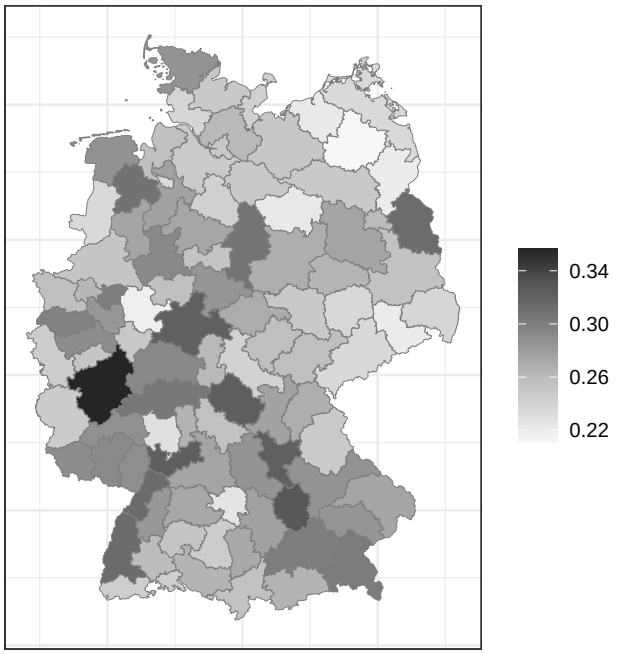


Figure 5.6: Model-based estimates of the Gini coefficients for SPRs.

5.6 Concluding remarks

Measuring inequality at a regionally detailed level within counties and municipalities can provide deep insight into the income and wealth structures of these entities and can serve policymakers to target policies, taxation and funding to address inequality. A common indicator for measuring inequality is the Gini coefficient, which can be applied equally to income before and after taxes or to the value of wealth. The approach presented provides model-based estimates of the Gini coefficients at a regionally detailed level, which entails a gain in precision for small sample sizes compared to direct estimates based only on survey data. To achieve this, additional data sources and information from other domains are used in addition to the survey data. As an alternative when micro-data is not available, an area-level model, namely a logit-transformed FH model, is applied to the nonlinear indicator of interest. To avoid a bias when transforming back from the logit scale to the original, a bias-corrected back-transformation is used, which is also incorporated into the parametric bootstrap to measure the uncertainty of the estimate. The methodology presented is a straightforward extension of elaborated results for the transformed FH-model, can be easily integrated into existing SAE software, such as the R-package **emdi** (Kreutzmann et al., 2019), and poses no computational challenges. The validity of the approach is demonstrated in a model-based simulation, where the point estimator also performs similarly well to the Bayesian approach of Fabrizi and Trivisano (2016) chosen for comparison. The methodology is illustrated by means of an example for German SPRs using survey data from the SOEP and data from the 2011 Census. The analysis shows that there are intra-regional differences in income inequality and the proposed model-based methodology has achieved the desired gain in precision. The approach can be readily applied to estimate Gini coefficients for other regions, sub-populations, or survey data.

For future research, the methodology could be extended to the use of survey data where the data have been imputed multiple times by the data provider due to item non-response. The approach of Kreutzmann et al. (2022), which uses multiply imputed data from the Household Finance and Consumption Survey to estimate wealth averages, could therefore be extended to nonlinear indicators and appropriate transformations to allow Rubin's pooling rules (Rubin, 1987) for multiply imputed data to be applied. Esteban et al. (2012) study area-level time models for nonlinear indicators such as poverty incidence and poverty gap. This approach could be transferred to also obtain time-stable estimates of inequality measures such as the Gini coefficient. Furthermore, the multivariate FH model proposed by Benavent and Morales (2016) could be extended for nonlinear indicators to jointly estimate Gini coefficients for multiple panel waves. Moreover, as mentioned in the introduction, other transformations could be used instead of the logit transformation as long as the estimated Gini coefficients are between zero and one. In any case, the variances of the direct estimator on the transformed scale are needed, and a suitable back-transformation for the estimated model-based Gini coefficients is required. Derivation of methodologies for e.g. probit or complementary log-log transformation could be part of further research.

Acknowledgements

The author appreciates gratefully the support of the German Research Foundation within the TESAP project (281573942). The data used in this publication were made available by the German SocioEconomic Panel Study (SOEP) at the German Institute for Economic Research (DIW), Berlin. The author is grateful for the computation time provided by the HPC service (`http://dx.doi.org/10.17169/refubium-26754`) of the Freie Universität Berlin.

Appendix E

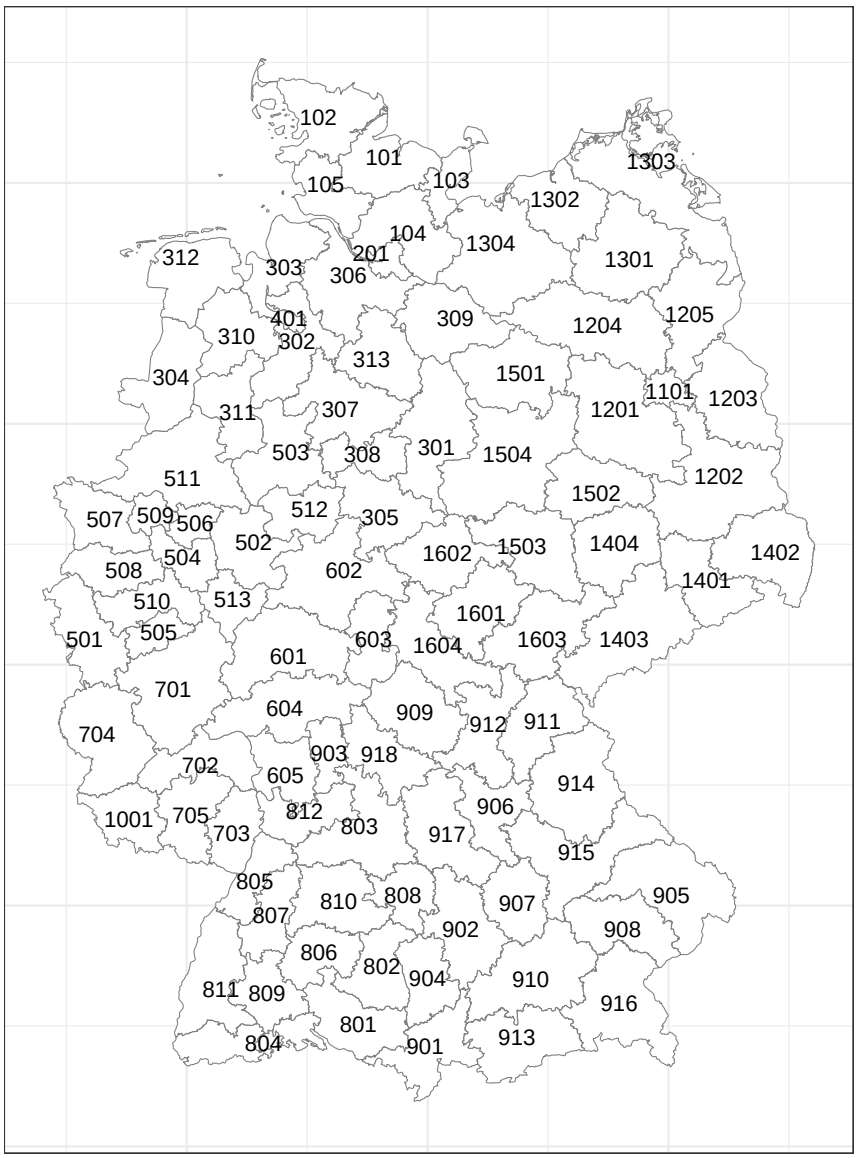


Figure E.1: SPR labels (Bundesinstitut für Bau-, Stadt-, und Raumforschung, 2017).

Table E.1: Official SPR names and labels (Bundesinstitut für Bau-, Stadt-, und Raumforschung, 2017).

SPR	Name	SPR	Name
101	Schleswig-Holstein Mitte	806	Neckar-Alb
102	Schleswig-Holstein Nord	807	Nordschwarzwald
103	Schleswig-Holstein Ost	808	Ostwürttemberg
104	Schleswig-Holstein Süd	809	Schwarzwald-Baar-Heuberg
105	Schleswig-Holstein Süd-West	810	Stuttgart
201	Hamburg	811	Südlicher Oberrhein
301	Braunschweig	812	Unterer Neckar
302	Bremen-Umland	901	Allgäu
303	Bremerhaven	902	Augsburg
304	Emsland	903	Bayerischer Untermain
305	Göttingen	904	Donau-Iller (BY)
306	Hamburg-Umland-Süd	905	Donau-Wald
307	Hannover	906	Industrieregion Mittelfranken
308	Hildesheim	907	Ingolstadt
309	Lüneburg	908	Landshut
310	Oldenburg	909	Main-Rhön
311	Osnabrück	910	München
312	Ost-Friesland	911	Oberfranken-Ost
313	Südheide	912	Oberfranken-West
401	Bremen	913	Oberland
501	Aachen	914	Oberpfalz-Nord
502	Arnsberg	915	Regensburg
503	Bielefeld	916	Südostoberbayern
504	Bochum/Hagen	917	Westmittelfranken
505	Bonn	918	Würzburg
506	Dortmund	1001	Saar
507	Duisburg/Essen	1101	Berlin
508	Düsseldorf	1201	Havelland-Fläming
509	Emscher-Lippe	1202	Lausitz-Spreewald
510	Köln	1203	Oderland-Spree
511	Münster	1204	Prignitz-Oberhavel
512	Paderborn	1205	Uckermark-Barnim
513	Siegen	1301	Mecklenburgische Seenplatte
601	Mittelhessen	1302	Mittleres Mecklenburg/Rostock
602	Nordhessen	1303	Vorpommern
603	Osthessen	1304	Westmecklenburg
604	Rhein-Main	1401	Oberes Elbtal/Osterzgebirge
605	Starkenburg	1402	Oberlausitz-Niederschlesien
701	Mittelrhein-Westerwald	1403	Südsachsen
702	Rheinhessen-Nahe	1404	Westsachsen
703	Rheinpfalz	1501	Altmark
704	Trier	1502	Anhalt-Bitterfeld-Wittenberg
705	Westpfalz	1503	Halle/S.
801	Bodensee-Oberschwaben	1504	Magdeburg
802	Donau-Iller (BW)	1601	Mittelthüringen
803	Franken	1602	Nordthüringen
804	Hochrhein-Bodensee	1603	Ostthüringen
805	Mittlerer Oberrhein	1604	Südthüringen

Table E.2: Direct variances on logit-scale of 89 observed SPRs.

SPR	$\sigma^2_{e_d}$	SPR	$\sigma^2_{e_d}$	SPR	$\sigma^2_{e_d}$	SPR	$\sigma^2_{e_d}$	SPR	$\sigma^2_{e_d}$
101	0.1949	504	0.1463	803	0.1778	911	0.1678	1404	0.1337
102	0.1036	505	0.1171	804	0.2092	912	0.4429	1501	0.2102
103	0.2641	506	0.1508	805	0.1317	914	0.1665	1502	0.1952
104	0.4305	507	0.1846	806	0.1240	915	0.3350	1503	0.1628
105	0.3098	508	0.0817	807	0.3318	916	0.1400	1504	0.1635
201	0.1679	509	0.1687	808	0.2026	917	0.2684	1601	0.1794
301	0.1987	510	0.1489	809	0.3849	918	0.2056	1602	0.4599
302	0.2580	511	0.1380	810	0.1172	1001	0.1711	1603	0.1075
304	0.1901	512	0.2584	811	0.1328	1101	0.0908	1604	0.2188
305	0.1387	601	0.2484	812	0.1767	1201	0.1984		
306	0.2301	602	0.1460	901	0.1668	1202	0.1365		
307	0.1931	603	0.5892	902	0.1524	1203	0.1515		
308	0.2093	604	0.0953	903	0.2411	1205	0.3145		
310	0.1808	605	0.1855	904	0.2236	1301	0.2276		
311	0.1887	701	0.1451	905	0.2047	1302	0.1896		
312	0.3857	702	0.2318	906	0.1310	1303	0.3853		
401	0.2335	703	0.1867	907	0.2520	1304	0.1725		
501	0.1741	704	0.2290	908	0.2093	1401	0.1997		
502	0.2215	705	0.1999	909	0.1703	1402	0.1657		
503	0.1334	802	0.3116	910	0.0906	1403	0.1383		

Table E.3: Distributions of MCEs of the ABias and RMSE values.

	Estimator	Min	1stQ	Median	Mean	3rdQ	Max
$10^3 \times \widehat{\text{MCE}}(\text{ABias})$	Bayesian Beta	0.456	0.726	0.820	0.815	0.916	1.097
	FH	0.480	0.734	0.849	0.841	0.926	1.155
	logit FH.bc	0.483	0.732	0.823	0.826	0.910	1.084
	logit FH.naive	0.483	0.733	0.830	0.827	0.910	1.083
$10^3 \times \widehat{\text{MCE}}(\text{RMSE})$	Bayesian Beta	0.329	0.526	0.598	0.591	0.664	0.793
	FH	0.326	0.523	0.603	0.600	0.664	0.932
	logit FH.bc	0.328	0.524	0.593	0.590	0.659	0.791
	logit FH.naive	0.329	0.525	0.595	0.592	0.662	0.794

Bibliography

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

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.

Alfons, A., M. Templ, and P. Filzmoser (2013). Robust estimation of economic indicators from survey samples based on pareto tail modeling. Journal of the Royal Statistical Society, Series C (Applied Statistics) 62, 271–286.

Altmann, K., R. Bernard, J. Le Blanc, E. Gabor-Toth, M. Hebbat, L. Kothmayr, T. Schmidt, P. Tzamourani, D. Werner, and J. Zhu (2020). The Panel on Household Finances (PHF)–microdata on household wealth in Germany. German Economic Review 21(3), 373–400.

Ammermüller, A., A. M. Weber, and P. Westerheide (2005). Die Entwicklung und Verteilung des Vermögens privater Haushalte unter besonderer Berücksichtigung des Produktivvermögens. Abschlussbericht zum Forschungsauftrag des Bundesministeriums für Gesundheit und Soziale Sicherung, Zentrum für Europäische Wirtschaftsforschung GmbH.

Arndt, O., H. Dalezios, P. Steden, and G. Färber (2009). Die regionale Inzidenz von Bundesmitteln. In H. Mäding (Ed.), Öffentliche Finanzströme und räumliche Entwicklung, pp. 9–48. Hannover: Verlag der Academy for Spatial Research and Planning (ARL).

Avila-Valdez, J. L., M. Huerta, V. Leiva, M. Riquelme, and L. Trujillo (2020). The Fay-Herriot model in small area estimation: EM algorithm and application to official data. REVSTAT-Statistical Journal 18, 613–635.

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 and Data Analysis 94, 372–390.

Biau, G. and E. Scornet (2016). A random forest guided tour. Test 25(2), 197–227.

Bijlsma, I., J. Brakel, R. Van der Velden, and J. Allen (2020). Estimating literacy levels at a detailed regional level: an application using Dutch data. Journal of Official Statistics 36, 251–274.

Bilton, P., G. Jones, S. Ganesh, and S. Haslett (2017). Classification trees for poverty mapping. Computational Statistics and Data Analysis 115, 53–66.

Blum, U., M. Brachert, H.-U. Brautzsch, K. Brenke, H. Buscher, D. Dietrich, W. Dürig, P. Franz, J. Günther, P. Haug, et al. (2011). Wirtschaftlicher Stand und Perspektiven für Ostdeutschland: Studie im Auftrag des Bundesministeriums des Innern. IWH-Sonderheft, Institut für Wirtschaftsforschung Halle.

Blum, U., H. S. Buscher, H. Gabrisch, J. Günther, G. Heimpold, C. Lang, U. Ludwig, M. T. W. Rosenfeld, and L. Schneider (2010). Ostdeutschlands Transformation seit 1990 im Spiegel wirtschaftlicher und sozialer Indikatoren. IWH-Sonderheft, Institut für Wirtschaftsforschung Halle.

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

Braml, M. and G. Felbermayr (2018). Regionale Ungleichheit in Deutschland und der EU: Was sagen die Daten? ifo Schnelldienst 71, 37–49.

Breiman, L. (1996). Bagging predictors. Machine Learning 24(2), 123–140.

Breiman, L. (2001). Random forests. Machine Learning 45(1), 5–32.

Brown, G., R. Chambers, P. Heady, and D. Heasman (2001). Evaluation of small area estimation methods: An application to unemployment estimates from the UK LFS. Symposium 2001 - Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Statistics Canada.

Budde, R. and L. Eilers (2014). Sozioökonomische Daten auf Rasterebene: Datenbeschreibung der microm-Rasterdaten. RWI Materialen, Leibniz-Institut für Wirtschaftsforschung.

bulwiengesa AG (2018). Intelligente Daten für klare Entscheidungen. `https://www.bulwiengesa.de/de/leistungsprogramm/daten`. [accessed: 09.2018].

Bundesinstitut für Bau-, Stadt-, und Raumforschung (2017). Indikatoren und Karten zur Raum- und Stadtentwicklung. `http://www.inkar.de/`. Datenlizenz Deutschland - Namensnennung - Version 2.0 [accessed: 04.2018/05.2022].

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.

Burnham, K. P. and D. R. Anderson (2010). Model selection and multimodel inference: a practical information-theoretic approach (2nd ed.). New York: Springer.

Butar, F. B. and P. Lahiri (2003). On measures of uncertainty of empirical bayes small-area estimators. Journal of Statistical Planning and Inference 112(1), 63–76.

Casas-Cordero, C., J. Encina, and P. Lahiri (2016). Poverty mapping for the Chilean comunas. In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation, pp. 379–404. Hoboken: John Wiley and Sons, Ltd.

Chambers, R., H. Chandra, N. Salvati, and N. Tzavidis (2014). Outlier robust small area estimation. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 76(1), 47–69.

Chandra, H., K. Aditya, and S. Kumar (2017). Small area estimation under a log transformed area level model. Journal of Statistical Theory and Practice 12, 497–505.

Chandra, H., N. Salvati, and R. Chambers (2015). A spatially nonstationary Fay-Herriot model for small area estimation. Journal of the Survey Statistics and Methodology 3(2), 109–135.

Council, N. R. (2000). Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology. Washington, DC: The National Academies Press.

Cowell, F. and E. Flachaire (2007). Income distribution and inequality measurement: The problem of extreme values. Journal of Econometrics 141, 1044–1072.

Da Maia, C. (2022, October). World bank group poverty & equity brief, Africa Eastern & Southern, Mozambique.

Datta, G. and P. Lahiri (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statistica Sinica 10(2), 613–627.

Datta, G. S., M. Ghosh, R. Steorts, and J. Maples (2011). Bayesian benchmarking with applications to small area estimation. Test 20(3), 574–588.

De Moliner, A. and C. Goga (2018). Sample-based estimation of mean electricity consumption curves for small domains. Survey Methodology 44(2). Statistics Canada.

Destatis (2018). Wirtschaftsrechnungen Einkommens- und Verbrauchsstichsprobe Einkommensverteilung in Deutschland 2013. Statistisches Bundesamt (Destatis).

Deutsche Bundesbank (2016). Vermögen und Finanzen privater Haushalte in Deutschland: Ergebnisse der Vermögensbefragung 2014. Monatsbericht, Deutsche Bundesbank.

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.

Edochie, I., D. Newhouse, T. Schmid, N. Tzavidis, E. Foster, A. Ouedraogo, A. Sanoh, and A. Savadogo (2023). Small area estimates of poverty in four West African countries. Working paper.

Eisele, M. and J. Zhu (2013). Multiple imputation in a complex household survey - the German panel on household finances (PHF): Challenges and solutions. User Guide, Deutsche Bundesbank.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). Micro level estimation of poverty and inequality. Econometrica 71, 355–364.

Empirica ag (2017). Regionaldatenbank Immobilien. `https://www.empirica-institut.de/`. [accessed: 10.2017].

Esteban, M., D. Morales, A. Perez, and L. Santamaria (2012). Small area estimation of poverty proportions under area-level time models. Computational Statistics and Data Analysis 56, 2840–2855.

European Commission (2017). Taxation and customs,taxes in europe database v3. `https://ec.europa.eu/taxation_customs/tedb/taxSearch.html`. [accessed: 08.02.2023].

Eurostat (2011a). Conventional dwellings by occupancy status, type of building and nuts 3 region. `http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=cens_11dwob_r3&lang=en`. [accessed: 04.2021].

Eurostat (2011b). Private households by type, tenure status and nuts 2 region. `http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=cens_11htts_r2&lang=en`. [accessed: 04.2021].

Eurostat (2017a). Harmonised unemployment rates. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ei_lmhr_m&lang=en. [accessed: 04.2021].

Eurostat (2017b). Main gdp aggregates per capita. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nama_10_pc&lang=en. [accessed: 04.2021].

Eurostat (2017c). Main national accounts tax aggregates. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=gov_10a_taxag&lang=en. [accessed: 04.2021].

Eurostat (2017d). Population: Structure indicators. http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_pjanind. [accessed: 04.2021].

Eurostat (2023). DataCollection: Precision Level DCF.

Fabrizi, E., M. R. Ferrante, and C. Trivisano (2016). Bayesian beta regression models for the estimation of poverty and inequality parameters in small areas. In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation, pp. 299–314. John Wiley and Sons, Ltd, Hoboken, NJ, USA.

Fabrizi, E. and C. Trivisano (2016). Small area estimation of the Gini concentration coefficient. Computational Statistics and Data Analysis 99, 223–234.

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

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

Fisher, J. (2006). Income imputation and the analysis of expenditure data in the consumer expenditure survey. U.S. Bureau of Labor Statistics, Working Papers.

Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014). Panel on Household Finances (PHF). Plus one additional attribute (district code).

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

Frick, J. R. and M. Grabka (2009). Gestiegene Vermögensungleichheit in Deutschland. DIW Wochenbericht, Deutsches Institut für Wirtschaftsforschung.

Furceri, D. and J. D. Ostry (2019). Robust determinants of income inequality. Oxford Review of Economic Policy 35(3), 490–517.

Gautier, E. (2011). Hierarchical Bayesian estimation of inequality measures with nonrectangular censored survey data with an application to wealth distribution of french households. The Annals of Applied Statistics 5(2), 1632–1656.

Gini, C. (1912). Variabilita emutabilita: contributo allo studio delle distribuzioni e delle relazioni statistiche. Econ. Giur. R. Univ. Cagliari 3, 3–159.

Giusti, C., S. Marchetti, M. Pratesi, and N. Salvati (2012). Semiparametric Fay-Herriot model using penalized splines. Journal of the Indian Society of Agricultural Statistics 1, 1–12.

Goebel, J. (2020). Informationen zu den SOEP-Geocodes (SOEP.V35). SOEP Survey Papers 407: Series D. Berlin: DIW/SOEP.

Goebel, J. and J. Frick (2005). Regional income stratification in unified Germany using a Gini decomposition approach. Regional Studies 42, 555–577.

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

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

Greenwell, B. M. (2017). **pdp**: An R package for constructing partial dependence plots. The R Journal 9(1), 421–436.

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, and A.-K. Kreutzmann (2020). Estimating regional unemployment with mobile network data for functional urban areas in Germany. `https://refubium.fu-berlin.de/handle/fub188/27030`.

Hagenaars, A., K. de Vos, and M. A. Zaidi (1994). Poverty Statistics in the Late 1980s: Research Based on Micro-data. Luxembourg: Office for Official Publications of the European Communities. Luxembourg.

Hajjem, A., F. Bellavance, and D. Larocque (2014). Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation 84, 1313–1328.

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

Hastie, T., R. Tibshirani, and J. Friedman (2008). The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer.

Herzog, A., S. V. Lall, J. Baez, P. Olinto, K. Simler, S. Nakamura, B. M. Zaengerling, H. Kim, D. S. Jones, B. P. Stewart, H. Cherkezian, L. Lima, B. W. M. V. Barros, T. Yepes, J. A. Oberreiter, and A. J. Acioly (2017, June). Greater Maputo: Urban poverty and inclusive growth (English). Working Papers World Bank Washington.

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

Household Finance and Consumption Network (2013a). The Household Finance and Consumption Survey: Methodological report for the first wave. Statistics Paper Series, European Central Bank.

Household Finance and Consumption Network (2013b). The Household Finance and Consumption Survey: Results from the first wave. Statistics Paper Series, European Central Bank.

Household Finance and Consumption Network (2016a). The Household Finance and Consumption Survey: Methodological report for the second wave. Statistics Paper Series, European Central Bank.

Household Finance and Consumption Network (2016b). The Household Finance and Consumption Survey: Results from the second wave. Statistics Paper Series, European Central Bank.

Household Finance and Consumption Network (2020a). The Household Finance and Consumption Survey: Results from the 2017 wave. Statistics Paper Series.

Household Finance and Consumption Network (2020b). The Household Finance and Consumption Survey: Wave 2017 statistical tables. Technical report.

Immel, L. and A. Peichl (2020). Regionale Ungleichheit in Deutschland: Wo leben die Reichen und wo die Armen? ifo Schnelldienst 73, 43–47.

Janicki, R. (2020). Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. Communications in Statistics-Theory and Methods 49, 2264–2284.

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Mathematica 30, 175–193.

Jiang, J. (2019). Robust Mixed Model Analysis. Hong Kong: World Scientific Publishing Company.

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

Jiang, J., P. Lahiri, and S.-M. Wan (2002). A unified jackknife theory for empirical best prediction with m-estimation. The Annals of Statistics 30(6), 1782–1810.

Jiang, J., P. Lahiri, S.-M. Wan, and C.-H. Wu (Eds.) (2001). Jackknifing in the Fay-Herriot model with an example. Proc. Sem. Funding Opportunity in Survey Research: Washington, DC: Bureau of Labor Statistics.

Jiang, J., T. Nguyen, and J. S. Rao (2011). Best predictive small area estimation. Journal of the American Statistical Association 106(494), 732–745.

Jiang, J. and J. Rao (2020). Robust small area estimation: An overview. Annual Review of Statistics and Its Application 7, 337–360.

Kara, S., S. Zimmermann, and SOEP Group (2019). SOEPcompanion (v34), v.2. SOEP Survey Papers 743: SeriesG. Berlin: DIW/SOEP.

Kim, J., J. Brick, W. Fuller, and G. Kalton (2006). On the bias of the multiple-imputation variance estimator in survey sampling. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 68(3), 509–521.

Knerr, P., F. Aust, N. Chudziak, R. Gilberg, and M. Kleudgen (2015). Methodenbericht - Private Haushalte und ihre Finanzen (PHF) 2. Erhebungswelle - Anonymisierte Fassung -. Methodenbericht, infas Institut für angewandte Sozialwissenschaft GmbH.

Koehler, E., E. Brown, and S. J.-P. A. Haneuse (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. The American Statistician 63(2), 155–162.

Kott, P. (1995). A paradox of multile imputation. Proceedings Survey Research Methods Section, American Statistical Association.

Krennmair, P. and T. Schmid (2022). Flexible domain prediction using mixed effects random forests. Journal of the Royal Statistical Society, Series C (Applied Statistics) 71, 1865–1894.

Kreutzmann, A.-K. (2019). Estimation of Disaggregated Indicators with Application to the Household Finance and Consumption Survey. Ph. D. thesis, Free University, Berlin.

Kreutzmann, A.-K., P. Marek, M. Runge, N. Salvati, and T. Schmid (2022). The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany. Journal of Applied Statistics 49(13), 3278–3299.

Kreutzmann, A.-K., P. Marek, N. Salvati, and T. Schmid (2019). Estimating regional wealth in Germany: How different are East and West really? Deutsche Bundesbank Discussion Paper 35/2019.

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, 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.

Lahiri, P. and J. N. K. Rao (1995). Robust estimation of mean squared error of small area estimators. Journal of the American Statistical Association 90(430), 758–766.

Lahiri, P. and J. B. Suntornchost (2015). Variable selection for linear mixed models with applications in small area estimation. The Indian Journal of Statistics 77(2), 312–320.

Landesamt für Statistik Niedersachsen (2014). Gebäude- und Wohnungsbestand in Deutschland - Erste Ergebnisse der Gebäude- und Wohnungszählung 2011. Technical report, Statistische Ämter des Bundes und der Länder.

Li, H. and P. Lahiri (2010). An adjusted maximum likelihood method for solving small area estimation problems. Journal of Multivariate Analysis 101(4), 882–892.

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

Liu, B., P. Lahiri, and G. Kalton (2014). Hierarchical bayes modeling of survey-weighted small area proportions. Survey methodology 40, 1–13.

Longford, N. (2004). Missing data and small area estimation in the UK labour force survey. Journal of the Royal Statistical Society, Series A (Statistics in Society) 167, 341–373.

Longford, N. T. (2005). Missing data and small-area estimation. London: Springer.

Maiti, T., H. Ren, and S. Sinha (2014). Prediction error of small area predictors shrinking both means and variances. Scandinavian Journal of Statistics 41, 775–790.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli (2015). Small area model-based estimators using big data sources. Journal of Official Statistics 31(2), 263–281.

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

Marshall, A., D. Altman, R. Holder, and P. Royston (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. BMC medical research methodology 9(57), 1–8.

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.

McConville, K. S. and D. Toth (2019). Automated selection of post-strata using a model-assisted regression tree estimator. Scandinavian Journal of Statistics 46(2), 389–413.

Mendez, G. (2008). Tree-based methods to model dependent data. Ph. D. thesis, Arizona State University. Unpublished doctoral dissertation.

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

Moura, F., A. Neves, and D. do N. Silva (2017). Small area models for skewed Brazilian business survey data. Journal of the Royal Statistical Society, Series A (Statistics in Society) 180(4), 1039–1055.

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

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. Estadistica 65(185), 13–37.

Newhouse, D., J. 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 Papers World Bank Washington (10175). License: CC BY 3.0 IGO.

OECD (2011). Income distribution database: by country - inequality. `https://stats.oecd.org`. [accessed: 06.2021].

OECD (2013). Regional well-being: Regional income distribution and poverty. `https://stats.oecd.org`. [accessed: 06.2021].

Opsomer, J., G. Claeskens, M. Ranalli, G. Kauermann, and F. Breidt (2008). Nonparametric small area estimation using penalized spline regression. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 70(1), 265–283.

Ottaviano, G. I. P. and D. Puga (1998). Agglomeration in the global economy: A survey of the 'new economic geography'. World Economy 21(6), 707–731.

Perugini, C. and G. Martino (2008). Income inequality within European regions: Determinants and effects on growth. Review of Income and Wealth 54, 373–406.

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

Piketty, T. and G. Zucman (2014). Capital is back: Wealth-income ratios in rich countries 1700-2010. The Quarterly Journal of Economics 129(3), 1255–1310.

Porter, A. T., S. H. Holan, C. K. Wikle, and N. Cressie (2014). Spatial Fay–Herriot models for small area estimation with functional covariates. Spatial Statistics 10, 27–42.

Prasad, N. and J. Rao (1990). The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association 85(409), 163–171.

Pratesi, M. (2016). Analysis of Poverty Data by Small Area Estimation. John Wiley and Sons, Inc, Hoboken, NJ, USA.

Rao, J. and I. Molina (2015). Small Area Estimation (2nd ed.). Hoboken: John Wiley and Sons, Inc, Hoboken, NJ, USA.

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

Rao, J. N. K. (1999). Some recent advances in model-based small area estimation. Survey Methodology 25(2), 175–186.

Rao, J. N. K. and C. F. J. Wu (1988). Resampling inference with complex survey data. Journal of the American Statistical Association 83(401), 231–241.

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

R Core Team (2022). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Riphahn, R. and O. Serfling (2005). Item non-response on income and wealth questions. Empirical Economics 30, 521–538.

Rivest, L.-P. and N. Vandal (2002). Mean squared error estimation for small areas when the small area variances are estimated. In J. Rao (Ed.), Proceedings of International Conference of Recent Advanced Survey Sampling, pp. 197–206. Ottawa, July 10-13, 2002. Available at: https://www.mat.ulaval.ca/fileadmin/mat/documents/lrivest/Publications/64-RivestVandal2003.pdf.

Rodríguez-Pose, A. and R. Crescenzi (2008). Mountains in a flat world: Why proximity still matters for the location of economic activity. Cambridge Journal of Regions, Economy and Society 1(3), 371–388.

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.

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.

Rubin, D. B. (1976). Inference and missing data. Biometrika 63(3), 581–592.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Hoboken: John Wiley and Sons, Hoboken.

Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Society 91(434), 473–489.

RWI;microm (2016a). Socio-Economic Data on grid level. Car segments. http://doi.org/10.7807/microm:pkwseg:v4. RWI-GEO-GRID. Version: 1. RWI - Leibniz-Institut für Wirtschaftsforschung. Datensatz.

RWI;microm (2016b). Socio-Economic Data on grid level. House typ. http://doi.org/10.7807/microm:haustyp:v4. RWI-GEO-GRID. Version: 1. RWI - Leibniz-Institut für Wirtschaftsforschung. Datensatz.

RWI;microm (2016c). Socio-Economic Data on grid level. Household structure. http://doi.org/10.7807/microm:hstruktur:v4. RWI-GEO-GRID. Version: 1. RWI - Leibniz-Institut für Wirtschaftsforschung. Datensatz.

RWI;microm (2016d). Socio-Economic Data on grid level. Payment index. http://doi.org/10.7807/microm:zahlindex:v4. RWI-GEO-GRID. Version: 1. RWI - Leibniz-Institut für Wirtschaftsforschung. Datensatz.

RWI;microm (2016e). Socio-Economic Data on grid level. population by age and gender. http://doi.org/10.7807/microm:einwGeAl:v4. RWI-GEO-GRID. Version: 1. RWI - Leibniz-Institut für Wirtschaftsforschung. Datensatz.

Santos, R. and V. Salvucci (2016). Poverty in Mozambique - significant progress but challenges remain. WIDER Policy Brief. Helsinki: UNU-WIDER.

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.

Schmid, T. and R. Münnich (2014). Spatial robust small area estimation. Statistical Papers 55, 653–670.

Schmid, T., N. Tzavidis, R. Münnich, and R. Chambers (2016). Outlier robust small area estimation under spatial correlation. Scandinavian Journal of Statistics 43(3), 806–826.

Seitz, W. (2019). Where they live: District-level measures of poverty, average consumption, and the middle class in Central Asia. Policy Research Working Papers World Bank Washington (8940). License: CC BY 3.0 IGO.

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

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

Siegers, R., V. Belcheva, and T. Silbermann (2020). SOEP-core v35 documentation of sample sizes and panel attrition in the German socio-economic panel (SOEP) (1984 until 2018). SOEP Survey Papers 826: Series C. Berlin: DIW/SOEP.

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

Slud, E. 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.

Socio-Economic Panel (2019). Data for years 1984-2017, version 34, SOEP. Socio-Economic Panel, Berlin. doi: 10.5684/soep.v34.

Statistische Ämter der Länder (2021). Volkswirtschaftliche Gesamtrechnungen der Länder - Zusammenhänge, Bedeutung, Ergebnisse. https://www.statistikportal.de/sites/default/files/2020-11/vgrdl_brochure_2020.pdf. [accessed: 06.2022].

Statistische Ämter des Bundes und der Länder (2011a). https://ergebnisse2011.zensus2022.de/datenbank/online/. [accessed: 01.2021].

Statistische Ämter des Bundes und der Länder (2011b). Erwerbstätige Bevölkerung im regionalen Vergleich nach Stellung im Beruf. https://ergebnisse.zensus2011.de/#StaticContent:00,BEG_4_3_2,,https://ergebnisse.zensus2011.de/#StaticContent:00,BEG_4_3_2,,. Zensus 2011 [accessed: 06.2018/06.2021].

Statistische Ämter des Bundes und der Länder (2011c). Regionaldatenbank Deutschland. https://www.regionalstatistik.de/genesis/online. [accessed: 01.2021].

Statistische Ämter des Bundes und der Länder (2014a). Arbeitslose nach ausgewählten Personengruppen sowie Arbeitslosenquoten - Jahresdurchschnitt - regionale Ebenen. https://www.regionalstatistik.de/genesis/online/data;jsessionid=303A27704A8955EAFF3BEDB689D51244.reg2?operation=abruftabelleBearbeiten&levelindex=2&levelid=1528188166412&auswahloperation=abruftabelleAuspraegungAuswaehlen&auswahlverzeichnis=ordnungsstruktur&auswahlziel=werteabruf&selectionname=13211-02-05-4-B&auswahltext=&nummer=10&variable=10&name=DLAND&werteabruf=Werteabruf. Regionaldatenbank Deutschland [accessed: 06.2018].

Statistische Ämter des Bundes und der Länder (2014b). Indikatoren des Indikatorensystems "Nachhaltigkeit" Themenbereich "Bevölkerung". https://www-genesis.destatis.de/gis/genView?GenMLURL=https://www-genesis.destatis.de/regatlas/AI-N-04.xml&CONTEXT=REGATLAS01. Regionalatlas Deutschland [accessed: 06.2018].

Statistische Ämter des Bundes und der Länder (2014c). Indikatoren des Themenbereichs "Bevölkerung". https://www-genesis.destatis.de/gis/genView?GenMLURL=https://www-genesis.destatis.de/regatlas/AI002-2.xml&CONTEXT=REGATLAS01. Regionalatlas Deutschland [accessed: 06.2021].

Statistische Ämter des Bundes und der Länder (2014d). Sparen der privaten Haushalte 1991 bis 2012 (WZ2008). https://www.statistik-bw.de/VGRdL/tbls/tab.jsp?rev=RV2011&tbl=tab15&lang=de-DE#tab05. Volkswirtschaftliche Gesamtrechnungen der Länder VGRdL [accessed: 06.2018].

Statistische Ämter des Bundes und der Länder (2014e). Verfügbares Einkommen 1991 bis 2016 (WZ2008). https://www.statistik-bw.de/VGRdL/tbls/tab.jsp?rev=RV2014&tbl=tab14&lang=de-DE#tab05. Volkswirtschaftliche Gesamtrechnungen der Länder VGRdL [accessed: 06.2018].

Statistische Ämter des Bundes und der Länder (2018). Gemeinsames Statistikportal. https://www.statistikportal.de/de/node/150. [accessed: 06.2018].

Steorts, R. C. and M. Ghosh (2013). On estimation of mean squared errors of benchmarked empirical bayes estimators. Statistica Sinica 23(2), 749–767.

Steorts, R. C., T. Schmid, and N. Tzavidis (2020). Smoothing and benchmarking for small area estimation. International Statistical Review 88, 580–598.

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

Sugasawa, S., H. Tamae, and T. Kubokawa (2017). Bayesian estimators for small area models shrinking both means and variances. Scandinavian Journal of Statistics 44, 150–167.

Sverchkov, M. and D. Pfeffermann (2018). Small area estimation under informative sampling and not missing at random non-response. Journal of the Royal Statistical Society, Series A (Statistics in Society) 181(4), 981–1008.

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 (2012). The future we want, outcome document of the united nations conference on sustainable development. Rio de Janeiro, Brazil, 20-22 June 2012.

United Nations Development Programme (2022). Human development report 2021/2022. https://hdr.undp.org/content/human-development-report-2021-22.

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

Van Buuren, S. (2018). Flexible Imputation of Missing Data Second Edition. Chapman and Hall/CRC.

Van Buuren, S. and K. Groothius-Oudshoorn (2011). **mice**: Multivariate imputation by chained equation in R. Journal of Statistical Software 45(3), 1–67.

Van Kerm, P. (2007). Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC. Working Paper Series 2007-01. Centre 39 dEtudes de Populations, de Pauvrete et de Politiques Socio-Economiques International Network for Studies in Technology, Environment, Alternatives, 40 Development.

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives 28(2), 3–28.

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

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

Wang, J. and W. A. Fuller (2003). The mean squared error of small area predictors constructed with estimated area variances. Journal of the American Statistical Association 98(463), 716–723.

Wright, M. N. and A. Ziegler (2017). **ranger**: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software 77, 1–17.

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.

Yoshimori, M. and P. Lahiri (2014). A new adjusted maximum likelihood method for the Fay-Herriot small area model. Journal of Multivariate Analysis 124, 281–294.

You, Y. and B. Chapman (2006). Small area estimation using area level models and estimated sampling variances. Survey Methodology 32(3), 97–103.

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

Zou, H. and T. Hastie (2005). Regularization and variable selection via the eslastic net. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 67(2), 301–320.

Summaries

Abstracts in English

Abstract: The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany

The increasing inequality of private income and wealth requires the redistribution of financial resources. Thus, several financial support schemes allocate budget across countries or regions. This work shows how to estimate private wealth at low regional levels by means of a modified Fay-Herriot approach that deals with (a) unit and item non-response, especially with used multiple imputation, (b) the skewness of the wealth distribution, and (c) inconsistencies of the regional estimates with the national direct estimate. One compelling example for financial redistribution is the promoted catching-up process of East Germany after the German reunification. This work shows that 25 years after the reunification differences are more diverse than just between the East and the West by estimating private wealth at two regional levels in Germany. The analysis is based on the Household Finance and Consumption Survey (HFCS) that the European Central Bank launched for all euro area countries in 2010. Although the application in this paper focuses particularly on Germany, the approach proposed is applicable to the other countries participating in the HFCS as well as to other surveys that make use of multiple imputation.
Keywords: Multiple imputation, Non-response, Small area estimation, Survey statistics

Abstract: Small area estimation with multiply imputed survey data

In this article, we propose a framework for small area estimation with multiply imputed survey data. Many statistical surveys suffer from (a) high nonresponse rates due to sensitive questions and response burden and (b) too small sample sizes to allow for reliable estimates on (unplanned) disaggregated levels due to budget constraints. One way to deal with missing values is to replace them by several plausible/imputed values based on a model. Small area estimation, such as the model by Fay and Herriot, is applied to estimate regionally disaggregated indicators when direct estimates are imprecise. The framework presented tackles simultaneously multiply imputed values and imprecise direct estimates. In particular, we extend the general class of transformed Fay-Herriot models to account for the additional uncertainty from multiple imputation. We derive three special cases of the Fay-Herriot model with particular transformations and provide point and mean squared error estimators. Depending on the case, the mean squared error is estimated by analytic solutions or resampling methods. Comprehensive simulations in a controlled environment show that the proposed methodology leads to reliable and precise results in terms of bias and mean squared error. The methodology is illustrated by a real data example using European wealth data.
Keywords: Fay-Herriot model, Mean squared error, Multiple imputation, Non-response, Survey statistics

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

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 step-wise 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

Abstract: Area-level small area estimation with random forests

This paper presents an approach that combines a small area estimation model with tree-based methods to provide a solution when only area-level data are available. In particular, the linear regression synthetic part of the Fay-Herriot model is replaced by a random forest to link survey data with related administrative information or data from other sources. By using a random forest, possible interactions among explanatory variables and nonlinear relationships between them and the dependent variable are accounted for. Automatic variable selection and robustness to outliers are indirectly provided as a property of the random forest. To obtain point estimates for a mean indicator, the familiar structure of the Fay-Herriot estimator is preserved. The estimation is done by implementing an expectation maximization algorithm. To determine the uncertainty of the point estimator, a nonparametric bootstrap method for estimating the mean squared error is presented. To evaluate the accuracy and precision of the proposed estimator and its uncertainty measure, model-based simulations are carried out. The presented methodology is illustrated by using household survey and remote sensing data from Mozambique to estimate average per capita consumption at a km grid-level.
Keywords: Fay-Herriot model, Remote sensing data, Survey statistics, Tree-based methods

Abstract: Estimating intra-regional inequality with an application to German spatial planning regions

Income inequality is a persistent topic of public and political debate. In this context, the focus often shifts from the national level to a more detailed geographical level. In particular, inequality between or within local communities can be assessed. In this paper, the estimation of inequality within regions, i.e. between households, is considered at a regionally dis-aggregated level. From a methodological point of view, a small area estimation of the Gini coefficient is carried out using an area-level model linking survey data with related administrative data. Specifically, the Fay-Herriot model is applied using a logit transformation followed by a bias-corrected back-transformation. The uncertainty of the point estimate is assessed using a parametric bootstrap procedure to estimate the mean squared error. The validity of the methodology is shown in a model-based simulation for the point estimator as well as for the uncertainty measure. The proposed methodology is illustrated by estimating model-based Gini coefficients for spatial planning regions in Germany, using survey data from the German Socio-Economic Panel and aggregate data from the 2011 Census. The results show that intra-regional inequality is more diverse than an east-west perspective would suggest.

Keywords: Fay-Herriot model, Gini coefficient, Small area estimation, Survey statistics

Kurzzusammenfassungen auf Deutsch

Zusammenfassung: Das Fay-Herriot-Modell für mehrfach imputierte Daten mit einer Anwendung auf regionale Vermögensschätzungen in Deutschland

Die steigende Ungleichheit der privaten Einkommen und Vermögen macht eine gerechte Umverteilung der finanziellen Ressourcen erforderlich. Dementsprechend werden im Rahmen verschiedener Finanzhilfeprogramme Haushaltsmittel zwischen Ländern oder Regionen umverteilt. In diesem Papier wird gezeigt, wie das private Vermögen auf einer kleineren regionalen Ebene mithilfe eines modifizierten Fay-Herriot-Ansatzes geschätzt werden kann, der folgende Aspekte berücksichtigt: a) Unit- und Item-Non-Response, insbesondere bei der Verwendung von multiplen Imputationen, b) die Schiefe der Vermögensverteilung, und c) Inkonsistenzen zwischen regionalen Schätzungen und nationalen direkten Schätzungen. Ein überzeugendes Beispiel für finanzielle Umverteilung ist der geförderte Aufholprozess Ostdeutschlands nach der deutschen Wiedervereinigung. Diese Arbeit zeigt, dass selbst 25 Jahre nach der Wiedervereinigung weiterhin erhebliche Unterschiede zwischen Ost und West bestehen, wenn das private Vermögen auf zwei regionalen Ebenen in Deutschland geschätzt wird. Die Analyse basiert auf dem Household Finance and Consumption Survey (HFCS), der im Jahr 2010 von der Europäischen Zentralbank für alle Länder der Eurozone eingeführt wurde. Obwohl sich die Anwendung in diesem Papier speziell auf Deutschland konzentriert, kann der vorgeschlagene Ansatz auch auf andere am HFCS teilnehmende Länder sowie auf andere Erhebungen angewendet werden, die Mehrfachimputationen verwenden.

Schlüssewörter: Mehrfach-Imputation, Non-Response, Small-Area-Schätzung, Survey-Statistik

Zusammenfassung: Small Area Schätzung mit mehrfach imputierten Erhebungsdaten

Viele statistische Erhebungen stehen vor zwei Herausforderungen: a) Hohe Antwortausfallraten aufgrund sensibler Fragen und eines aufwändigen Beantwortungsprozesses, sowie b) Budgetbeschränkungen, die zu kleinen Stichproben führen und somit keine zuverlässigen Schätzungen auf disaggregierten Ebenen ermöglichen. Ein Lösungsansatz für den Umgang mit fehlenden Werten besteht darin, diese durch mehrere plausible/imputierte Werte zu ersetzen, die auf einem Modell basieren. Small Area Modelle, wie das Fay-Herriot-Modell, werden verwendet, um regional disaggregierte Indikatoren zu schätzen, wenn direkte Schätzungen aufgrund kleiner Stichprobenumfänge ungenau sind. In diesem Papier schlagen wir einen Ansatz vor, der beide Probleme gleichzeitig angeht. Konkret erweitern wir die allgemeine Modellklasse der transformierten Fay-Herriot-Modelle, um die zusätzliche Unsicherheit durch multiple Imputationen zu berücksichtigen. Wir leiten drei Spezialfälle des Fay-Herriot-Modells mit spezifischen Transformationen ab und liefern Punkt- und mittlere quadratische Fehlerschätzer. Je nach Fall wird der mittlere quadratische Fehler entweder durch analytische Lösungen oder Resampling-Methoden geschätzt. Um die Zuverlässigkeit und Genauigkeit der vorgeschlagenen Methodik zu überprüfen, führen wir umfangreiche Simulationen in einer kontrollierten Umgebung durch. Die Ergebnisse zeigen, dass unsere Methodik verlässliche und präzise Schätzungen bezüglich Verzerrung und mittlerem quadratischem Fehler liefert. Um die Anwendung der Methode in der Praxis zu demonstrieren, verwenden wir ein reales Datenbeispiel mit europäischen Vermögensdaten.

Schlüssewörter: Fay-Herriot-Modell, Mittlerer quadratischer Fehler, Mehrfach-Imputation, Non-Response, Survey-Statistik

Zusammenfassung: Variablenselektion mit konditionalem AIC für lineare gemischte Modelle unter Verwendung datengetriebener Transformationen

Bei der Verwendung linearer gemischter Modelle stoßen Datenanalysten üblicherweise auf zwei praktische Probleme: a) Das wahre Modell ist unbekannt, und b) die Annahmen über die Fehlerterme nach dem Gaußschen Modell sind nicht gültig. Obwohl diese Probleme oft zusammen auftreten, behandeln Wissenschaftler sie tendenziell separat. Sie versuchen a) ein optimales Modell auf der Grundlage des bedingten Akaike-Informationskriteriums (*cAIC*) zu finden und b) Transformationen auf die abhängige Variable anzuwenden. Jedoch hängt das optimale Modell von der Transformation ab und umgekehrt. In diesem Papier haben wir das Ziel, beide Probleme gleichzeitig anzugehen. Insbesondere schlagen wir eine angepasste Form des *cAIC* vor, bei der die Jacobian der jeweiligen Transformation einbezogen wird, um verschiedene Modellkandidaten mit unterschiedlich transformierten Daten vergleichen zu können. Um dies numerisch zu bewältigen, stellen wir ein schrittweises Auswahlverfahren vor, das auf dem eingeführten angepassten *cAIC* basiert. Wir nutzen modellbasierte Simulationen, um das vorgeschlagene Auswahlverfahren mit alternativen Methoden zu vergleichen. Schließlich wenden wir den vorgestellten Ansatz auf mexikanische Daten an, um Armuts- und Ungleichheitsindikatoren für 81 Gemeinden zu schätzen und wollen dadurch demonstrieren, wie unser Ansatz in der Praxis angewendet werden kann.
Schlüssewörter: Box-Cox-Transformation, Empirischer bester Prädiktor, Indikatoren, Small-Area-Schätzung

Zusammenfassung: Kleinräumige Schätzung mit Random Forests für Daten auf Gebietsebene

Dieses Papier präsentiert einen innovativen Ansatz, der ein kleinräumiges Schätzmodell mit baumbasierten Methoden kombiniert, wenn lediglich Daten auf Gebietsebene vorliegen. Speziell wird der synthetische Teil der linearen Regressionskomponente des Fay-Herriot-Modells durch einen Random Forest ersetzt, um Erhebungsdaten mit entsprechenden Zensus-/Registerdaten oder anderen Quellen zu verknüpfen. Der Einsatz eines Random Forest ermöglicht die Berücksichtigung möglicher Wechselwirkungen zwischen den erklärenden Variablen sowie nichtlinearer Beziehungen zwischen ihnen und der abhängigen Variable. Darüber hinaus umfasst der Random Forest automatische Variablenselektion und Robustheit gegenüber Ausreißern als implizite Eigenschaften. Die Punktschätzungen für einen Mittelwertindikator werden unter Beibehaltung der bekannten Struktur des Fay-Herriot-Schätzers erhalten. Die Schätzung erfolgt mithilfe eines Erwartungsmaximierungsalgorithmus. Um die Unsicherheit des Punktschätzers zu bestimmen, wird ein nichtparametrisches Bootstrap-Verfahren zur Schätzung des mittleren quadratischen Fehlers eingeführt. Modellbasierte Simulationen werden durchgeführt, um die Genauigkeit und Präzision des vorgeschlagenen Schätzers und seines Unsicherheitsmaßes zu evaluieren. Zur Veranschaulichung der Methodik wird diese anhand von Haushaltserhebungen und Fernerkundungsdaten aus Mosambik angewendet, um den durchschnittlichen Pro-Kopf-Verbrauch auf einer Ebene von km-Rastern zu schätzen.
Schlüssewörter: Baum-basierte Methoden, Fay-Herriot model, Fernerkundungsdaten, Survey-Statistik

Zusammenfassung: Schätzung intraregionaler Ungleichheit mit einer Anwendung auf deutsche Raumordnungsregionen

Einkommensungleichheit ist ein fortlaufendes Thema in der öffentlichen und politischen Debatte. Dabei verschiebt sich der Fokus oft von nationalen Betrachtungen hin zu einer detaillierteren geografischen Ebene, um die Ungleichheit zwischen oder innerhalb lokaler Gemeinschaften zu untersuchen. In diesem Beitrag liegt der Fokus auf der Schätzung der Ungleichheit innerhalb von Regionen, insbesondere zwischen Haushalten, auf einer regional disaggregierten Ebene.´ Methodisch erfolgt die kleinräumige

Schätzung des Gini-Koeffizienten durch die Verknüpfung von Umfragedaten mit entsprechenden Verwaltungsdaten auf regionaler Ebene. Dazu wird das Fay-Herriot-Modell mit einer Logit-Transformation und einer anschließenden verzerrungskorrigierten Rücktransformation verwendet. Die Unsicherheit der Punktschätzung wird mittels eines parametrischen Bootstrap-Verfahrens zur Schätzung des mittleren quadratischen Fehlers bewertet. Die Gültigkeit dieser Methodik wird in einer modellbasierten Simulation sowohl für den Punktschätzer als auch für das Unsicherheitsmaß gezeigt. Zur Veranschaulichung der Methodik werden modellbasierte Gini-Koeffizienten für Raumordnungsregionen in Deutschland geschätzt. Dabei werden Befragungsdaten des Sozio-oekonomischen Panels und Aggregatdaten des Zensus 2011 verwendet. Die Ergebnisse verdeutlichen, dass die intraregionalen Disparitäten viel facettenreicher sind, als es eine einfache Ost-West-Perspektive vermuten lässt.
Schlüssewörter: Fay-Herriot-Modell, Gini-Koeffizient, Small-Area-Schätzung, Survey-Statistik

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

Berlin, February 27, 2023

Marina Runge
February 27, 2023