Thesis submitted in fulfilment of the requirements for the degree

Dr. rer. pol.

on the topic

Small Area Estimation under Limited Auxiliary Population Data Dealing with Model Violations and their Economic Applications

to the

Chair of Applied Statistics
School of Business and Economics
Freie Universität Berlin

submitted by

Nora Würz
born in Fulda

Berlin, 2022

Nora Würz, *Small Area Estimation under Limited Auxiliary Population Data Dealing with Model Violations and their Economic Applications,* July 2022

Supervisors:
Prof. Dr. Timo Schmid (Otto-Friedrich-Universität Bamberg)
Prof. Nikos Tzavidis, Ph.D. (University of Southampton)

Location:
Berlin

Date of defense:
December 19, 2022

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Dr. Timo Schmid (Otto-Friedrich-Universität Bamberg, Germany). His guidance, encouragement, mentoring as well as fruitful discussions have been invaluable for the success of this project.

I am also very thankful to Prof. Nikos Tzavidis, Ph.D. (Southampton University, England), for his input and ideas, our valuable and interesting discussions and his important suggestions.

Special thanks go to Studienstiftung des Deutschen Volkes for supporting this thesis by a scholarship.
This thesis has also been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 730998 (InGRID-2 - Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy).

Furthermore, I am very grateful to all others who have accompanied me on the way to this thesis, especially my colleagues at the Chair of Statistics and the Statistical Consulting Unit *fu:stat* for the pleasant and supporting working environment and friendship.

Last but not least, I want to thank my beloved family, especially my husband Jan for his constant support and my daughter Ida for giving me the distraction I needed.

Publication List

The publications listed below are the result of the research carried out in this thesis titled, "Small Area Estimation under Limited Auxiliary Population Data Dealing with Model Violations and their Economic Applications."

1. Würz, N., Schmid, T., and Tzavidis, N. (2022) **Estimating regional income indicators under transformations and access to limited population auxiliary information**, *Journal of the Royal Statistical Society: Series A (Statistics in Society), 185(4), pp. 1679-1706,* doi: `https://doi.org/10.1111/rssa.12913`. Accepted and published.

2. Würz, N. (2022) **The R package saeTrafo for estimating unit-level small area models under transformations.** *R package vignette*, `https://CRAN.R-project.org/package=saeTrafo`. Accepted and published.

3. Krennmair, P., Würz, N., and Schmid, T. (2022) **Analysing opportunity cost of care work using mixed effects random forests under aggregated census data.** *Working paper*, to be submitted. Preliminary work is available in Krennmair et al. (2022).

4. Dawber, J., Würz, N., Smith, P., Flower, T., Thomas, H., Schmid, T., and Tzavidis, N. (2022) **Experimental UK regional consumer price inflation with model-based expenditure weights.** *Journal of Official Statistics, 38(1), pp. 213-237,* doi: `https://doi.org/10.2478/jos-2022-0010`. Accepted and published.

5. Hadam, S., Würz, N., Kreutzmann, A.-K., and Schmid, T. (2022) **Estimating regional unemployment with mobile network data for functional urban areas in Germany,** resubmitted after major revision. The work is an extension of Hadam et al. (2020).

Contents

Introduction

For evidence-based policy-making, reliable information on socio-economic indicators are essential. Sample surveys have a long tradition of providing cost-efficient information on these indicators. Mostly, there is a demand for the quantity of interest not only at the level of the total population, but especially at the level of sub-populations (geographic areas or socio-demographic groups) called areas or domains. To gain insights into these sub-populations, disaggregated direct estimators can be used, which are calculated solely on area-specific survey data. An area is regarded as 'large' if the sample size is large enough to enable reliable direct estimates. If the precision of the direct estimates is not sufficient or the sample size is even zero, the area is considered as 'small'. This is particularly common at high spatial or socio-demographic resolutions. Small area estimation (SAE) is promising to overcome this problem without the need for larger and thus more costly surveys (Pfeffermann, 2013; Rao and Molina, 2015; Tzavidis et al., 2018). The essence of SAE techniques is that they 'borrow strength' from other areas to improve their predictions. For this purpose, a model is built on survey data that links additional auxiliary data and exploits area-specific structures. Suitable auxiliary data sources are administrative and register data, such as the census. In many countries, such data are strictly protected by confidentiality agreements and access to population micro-data is a challenge even for gatekeeper organisations. Thus, users have an increased interest in SAE estimators that do not require population micro-data to serve as auxiliary data. In this thesis, new methods in the absence of population micro-data are presented and applications on socio-economic highly relevant indicators are demonstrated.

Since different SAE models impose different data requirements, Part I bundles research combining unit-level survey data and limited auxiliary data, e.g., aggregated data such as means, which is a common data situation for users. To account for the unit-level survey information the use of the well-known nested error regression (NER) model from Battese et al. (1988) is targeted. This model is a special case of a linear mixed model based on several assumptions. But how can users proceed if the model assumptions are not fulfilled? In Part I, this thesis provides two new approaches to deal with this issue. One promising approach is to transform the response. Since several socio-economically relevant variables, such as income, have a skewed distribution, the log-transformation of the response is an established way to meet the assumptions (Berg and Chandra, 2014; Molina and Martín, 2018). However, the data-driven log-shift transformation is even more promising because it extends the log by an additional parameter and achieves more flexibility (Sugasawa and Kubokawa, 2019; Rojas-Perilla et al., 2020). Chapter 1 introduces both transformations in the absence of population micro-data. A particular challenge is the transformation of the small area means back to the

original scale. Hence, the proposed approach introduces aggregate statistics (means and covariances) and kernel density estimation to resolve the issue of lacking population micro-data. Uncertainty estimation is developed, and all methods are evaluated in design- and model-based settings. The proposed method is applied to estimate regional income in Germany using the Socio-Economic Panel (Socio-Economic Panel, 2019) and census data (Statistisches Bundesamt, 2015). It achieves a clear improvement in reliability, and thus demonstrates the importance of the method. To conveniently enable further applications, this new methodology is implemented in the R package **saeTrafo** (R Core Team, 2022; Würz, 2022). Chapter 2 describes the various functionalities of the package using publicly available income data. To increase user-friendliness, established unit-level models under transformations and their uncertainty estimations are implemented and the most suitable method is automatically selected. For some applications, however, it is challenging to find a suitable transformation or, more generally, to specify a model, particularly in the presence of complex interactions. For this case, machine learning methods are valuable as a transformation is not necessarily required nor a model needs to be explicitly specified (Hastie et al., 2009; Varian, 2014). The semi-parametric framework of mixed effects random forest (MERF) combines the advantages of random forests (robustness against outliers and implicit model-selection) with the ability to model hierarchical dependencies as present in SAE approaches (Krennmair and Schmid, 2022). Chapter 3 introduces MERFs in the absence of population micro-data. As existing random forest algorithm require unit-level auxiliary population data, an alternative strategy is introduced. It adaptively incorporates aggregated auxiliary information through calibration-weights to circumvent unit-level auxiliary data. Applying the proposed method on opportunity costs of care work for Germany using the Socio-Economic Panel (Socio-Economic Panel, 2019) and census data (Statistisches Bundesamt, 2015) demonstrates the gain in accuracy in comparison to both direct estimates and the classical NER model.

In contrast to methods using a unit-level sample survey, Part II focuses on the well-known class of area-level SAE models (Fay and Herriot, 1979) requiring direct estimates from a survey while using (once again) only aggregated population auxiliary data. This thesis presents two particularly relevant applications of this model class. Chapter 4 examines regional consumer price indices (CPIs) in the United Kingdom (UK), contributing to the great interest in monitoring inflation at the spatial level (Fenwick and O'Donoghue, 2003). The SAE challenge is to construct model-based expenditure weights to generate the regional basket of goods and services for the twelve regions of the UK. They are estimated and constructed from the living cost and food survey (Defra and ONS, 2019). Furthermore, available price data (ONS, 2020) are linked to the SAE estimated baskets to produce regional CPIs. The resulting CPI series are closely examined, and smoothing techniques are applied. As a result, the reliability improves, but the CPI series are still too volatile for policy use. However, our research serves as a valuable framework for the creation of a regional CPI in the future. The second application also explores the reliability of the disaggregated estimation of a politically and economically highly relevant indicator, in this case the unemployment rate. The regional target level are the functional urban areas in the German federal state North Rhine-Westphalia. In Chapter 5, two types of unemployment rates - the traditional one and an alternative definition taking com-

muting into account (Grözinger, 2018) - are estimated and compared. Direct estimates from the labour force survey (Eurostat, 2019b) are linked with SAE methods to passively collected mobile network data. This alternative data source is real-time available, offers spatial flexible resolutions, and is dynamic (Toole et al., 2015; Marchetti et al., 2015; Steele et al., 2017; Schmid et al., 2017). In compliance with data protection rules, we obtain aggregated auxiliary mobile network information from the data provider. The SAE methods improve the reliability, and the resulting predictions show that alternative unemployment rates in German city cores are lower than traditional estimated official unemployment rates indicate.

Part I

Strategies to Deal with Model Violations in Unit-Level Models under Limited Auxiliary Population Data

Chapter 1

Estimating regional income indicators under transformations and access to limited population auxiliary information

Chapter 2

The R package saeTrafo for estimating unit-level small area models under transformations

2.1 Introduction

For evidence-based policymaking, reliable knowledge of the spatial distribution of important variables like income is essential. As sample sizes are small at a high-resolution spatial scale of interest, direct estimates from surveys at this scale are likely to be unreliable. Small area estimation (SAE) methods are a promising and widely used approach to overcome this problem (Pfeffermann, 2013; Rao and Molina, 2015; Tzavidis et al., 2018). One predominant approach - for estimating the averages in small areas - is the nested error regression (NER) model proposed by Battese et al. (1988) that borrows strength by using auxiliary information from a census. The starting point for this model is the availability of survey data at the individual-level. For the census data, aggregates at the spatial scale of interest are sufficient. As small area models often rely on linear mixed models, the normality assumption for the error terms has to be satisfied. However, in a variety of real-world examples, this assumption is hard to meet. Especially skewed variables, like income and consumption, can often not be adequately described by the available auxiliary variables and lead to error terms where normality assumptions are rejected. One promising approach satisfying the assumptions of the NER model is to use fixed logarithmic (Molina and Martín, 2018) or data-driven (Sugasawa and Kubokawa, 2019; Rojas-Perilla et al., 2020) transformations for the dependent variable. When a back-transformation to the original scale is needed, a general problem is the bias-correction. Berg and Chandra (2014) suggest an estimator with minimal mean squared error (MSE). For this estimator, Molina and Martín (2018) develop an analytical MSE estimator. It requires auxiliary information from population micro-data to correct the bias caused by the back-transformation, which is a strong limitation for data analysts. Especially in countries with high data confidentiality standards, access to individual data from the census is usually not possible. For this need, Würz et al. (2022) proposed methodology for estimating small area means based on the transformed NER model, if only aggregate population-level auxiliary information is available. Their approach

presents an appropriate bias-correction that is necessary due to the back-transformation in the absence of population micro-data. It abstains from any parametric assumptions about the auxiliary variables and instead uses aggregate statistics (means and covariances) and kernel density estimation (KDE) to resolve the issue of not having access to population micro-data. The authors introduce a parametric bootstrap MSE estimator that captures the uncertainty caused by the use of transformations and KDE. Alternatively, Li et al. (2019) propose another method relying on the smearing approach of Duan (1983) but without introducing an MSE estimator. For the second major class of small area models for estimating means - the area-level models (Fay and Herriot, 1979) - aggregated survey and population data are sufficient to determine small area means. In addition, considerable research has been done for area-level models on the application of transformations: Slud and Maiti (2006) present an estimator for small area means and its analytical MSE estimator under a log transformed Fay-Herriot model. Sugasawa and Kubokawa (2017) discuss area-level models for the data-driven dual power transformations. However, this model class only employs aggregates from survey data. If the user has access to individual survey data, it would be desirable to account for this finer level of survey information by applying unit-level models.

For the estimation of small area means and indicators, several software packages exist. In the following, the R software packages (R Core Team, 2022) for estimating unit-level SAE models are briefly described: the package **rsae** (Schoch, 2014) focuses on robust estimation for both unit- and area-level SAE models but do not offer transformations. Both models are also available in the R package **JoSAE** (Breidenbach, 2018) or **rhnerm** (Sugasawa, 2016). They focus on the estimation under heteroscedasticity. The R package **hbsae** (Boonstra, 2022) fits both models by maximum likelihood or hierarchical Bayesian approaches. Like the previously listed R packages, **mcmcsae** (Boonstra, 2021) also does not provide the possibility for the use of transformations. It deals with correlated random effects for both unit- and area-level models and uses markov chain monte carlo simulations. The R package **sae** (Molina and Marhuenda, 2015) offers unit-level models together with a variety of area-level models. On the one hand, it provides the classic NER model (function: `eblupBHF`). On the other hand, a NER model with transformations (box-cox and power transformation (Box and Cox, 1964)) is available, but micro-population auxiliary data is required (function: `ebBHF`). For both models, bootstrap MSEs are available. However, it is important to emphasise that `ebBHF` requires population micro-data, which is a strict limitation for data analysts. A package providing transformations for SAE methods is the **emdi** package (Kreutzmann et al., 2019). It offers the area-level model and the method of Molina and Rao (2010), which requires individual census data.

The structure of **saeTrafo** is closely oriented on that of the R package **emdi** (Kreutzmann et al., 2019). This means that **saeTrafo** offers similar input arguments and generic functions. The main focus of **saeTrafo** lies on making the new methodology by Würz et al. (2022) publicly available to enable the use of transformations (log transformation and data-driven log-shift transformation) under limited auxiliary data for unit-level small area models. The relevance is justified by data confidentiality because in developed countries like Germany, population micro-data are not publicly available, and access to such data is even challenging within gatekeeper organizations. Instead, population-level auxiliary data are often only available at some

aggregate level. Furthermore, the use of transformations is essential to meet the assumptions on the error terms. Additionally, **saeTrafo** offers further methodology in a user-friendly way: the well-known model from Battese et al. (1988) (without transformations), the bias-corrected estimator from Molina and Martín (2018) (which requires population micro-data), and a first-order bias-corrected estimator in the presence of aggregated population data. Depending on the used data and transformation **saeTrafo** automatically selects the appropriate method. Furthermore, the user benefits from the simple determination of the uncertainty via the main function. Some uncertainty estimates rely on bootstrap procedures. For that, **saeTrafo** supplies a parallelization option to reduce running time. Moreover, it offers well-known and SAE-specific generic functions enabling the automatic generation of plots for model diagnostics, the comparison to a direct estimator via plots, the visualization of the estimates on a map, and the easy export of the results. As the relevant graphics are generated directly within the package and personalisation options exist, it simplifies the work flow for the user.

The rest of the paper is structured as follows: Section 2.2 introduces the estimation methods. In Section 2.3, the Austrian dataset which is used to illustrate the package is described. The functionalities of **saeTrafo** are presented in Section 2.4. This section gives a general overview on the main function `NER_Trafo`, demonstrate this function on exemplary Austrian data, and presents generic functions for the corresponding S3 object. Section 2.5 outlines further potential extensions.

2.2 Statistical methods

The package **saeTrafo** focuses on the NER model of Battese et al. (1988), which uses unit-level sample data and aggregated population-level auxiliary information. For a general overview on SAE, we refer to Rao and Molina (2015) or Tzavidis et al. (2018). This section presents the theoretical background starting from the classical NER model to the methodology from Würz et al. (2022).

2.2.1 The nested error regression model

Throughout the paper, a finite population U of size N is divided into D areas $U_1, U_2, ..., U_D$ consisting of $N_1, N_2, ..., N_D$ units. The index $i = 1, ..., D$ indicates the respective area and $j = 1, ..., N_i$ the corresponding units. The response y_{ij} is available for every unit in the sample s which consists of n units partitioned into sample sizes $n_1, n_2, ..., n_D$ for each area. With s_i / $\overline{s}_i$ we refer to the in-sample/out-of-sample units in area i. The vector $\mathbf{x}_{ij} = (1, x_1, x_2, ..., x_p)^T$ contains the intercept and p explanatory variables for every unit j in the sample. These vectors are combined within the matrix $\mathbf{X}_s$. The vector $\mathbf{y}_s$ contains the response of the individuals within the sample. The NER model of Battese et al. (1988) models the relationship between $\mathbf{x}_{ij}$ and y_{ij} as follows:

$$y_{ij} = \mathbf{x}_{ij}^T \beta + u_i + e_{ij}, \qquad u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2) \text{ and } e_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2), \tag{2.1}$$

where $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ is the vector of regression coefficients. u_i denotes the area-specific random effect and e_{ij} is the unit-level error. They are assumed to be independent and σ_u^2 and σ_e^2 denote their variances. An out-of-sample unit is estimated as best linear unbiased prediction by $\hat{\mu}_{ij} = \mathbf{x}_{ij}^T\hat{\beta} + \hat{u}_i = \mathbf{x}_{ij}^T\hat{\beta} + \hat{\gamma}_i \left(\sum_{j\in s_i}\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\beta}\right)\right)$, where $\hat{\gamma}_i = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/n_i}$ denotes the estimated shrinkage factor. The target parameter is the population mean for each area i and it is estimated as the empirical best linear unbiased predictor (EBLUP) for the population area mean ($\overline{y}_i$) by

$$\begin{aligned}\hat{\overline{Y}}_i^{\text{BHF}} &= \frac{1}{N_i}\left(\sum\nolimits_{j\in s_i} y_{ij} + \sum\nolimits_{j\in \overline{s}_i} \hat{\mu}_{ij}\right) \\ &= \hat{\gamma}_i \left(\frac{1}{n_i}\sum_{j\in s_i} y_{ij} + \left(\overline{\mathbf{x}}_i - \frac{1}{n_i}\sum_{j\in s_i}\mathbf{x}_{ij}\right)^T \hat{\beta}\right) + (1-\hat{\gamma}_i)\,\overline{\mathbf{x}}_i^T\hat{\beta}. \end{aligned} \tag{2.2}$$

The vector $\overline{\mathbf{x}}_i^T = \frac{1}{N_i}\sum_{j\in U_i}\mathbf{x}_{ij}^T$ contains means for the p covariates within i. **saeTrafo** uses the restricted maximum likelihood (REML) theory to estimate fixed effects and the variance components. As in the package **emdi** (Kreutzmann et al., 2019), it is implemented based on the `lme` function of the package **nlme** (Pinheiro et al., 2022). Note that the estimator of Battese et al. (1988) ($\hat{\overline{Y}}_i^{\text{BHF}}$, (2.2)) requires only population-level aggregates and a unit-level survey.

To estimate the uncertainty of $\hat{\overline{Y}}_i^{\text{BHF}}$ (2.2), Prasad and Rao (1990) propose an analytical MSE which **saeTrafo** supplies. A second possibility for determining the uncertainty are bootstrap methods offered by R packages such as **sae** (Molina and Marhuenda, 2015).

2.2.2 Small area estimation under the nested error regression model and transformations

One-to-one transformations of the response $h(y_{ij}) = y_{ij}^*$ are a common tool to prevent violations of the model assumptions. For skewed variables, like income, this problem is typical. In order to adapt better to the data, data-driven transformations are promising for SAE (Gurka et al., 2006; Rojas-Perilla et al., 2020). For instance, the log-shift transformation (Yang, 1995) extends the log transformation by including a transformation parameter λ: $y_{ij}^* = h(y_{ij}) = \log(y_{ij} + \lambda)$, which is estimated from the sample. In **saeTrafo**, the transformation parameter λ is estimated from the sample data using the REML method as Rojas-Perilla et al. (2020) proposed.

Using a transformation on the response results in a model on the transformed scale:

$$h(y_{ij}) = y_{ij}^* = \mathbf{x}_{ij}^T\beta + u_i + e_{ij}, \qquad u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2) \text{ and } e_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2). \tag{2.3}$$

The BLUP on the transformed scale for out-of-sample units is $\mu_{ij}^* = \mathbf{x}_{ij}^T\beta + u_i$. However, in SAE applications there is interest in prediction, so the aim is to estimate the mean on the original scale.

Due to Jensen's inequality, the naive back-transformation of real convex or concave functions $h()$ don't lead to the same result as the best prediction on the original scale (Jensen et al.,

1906):

$$\underbrace{\mu_{ij}^{\text{trans, naive}} = h^{-1}\left(\mu_{ij}^{*}\right)}_{\text{naive back-transformation of the BLUP}} \neq \underbrace{E[h^{-1}(y_{ij}^{*})|\mathbf{y}_s, \mathbf{X}_s]}_{\text{best prediction on original scale}}.$$

For the log and log-shift transformation, the back-transformation $h^{-1}() = \exp()$ or $h^{-1}() = \exp() - \lambda$ is convex and hence $\mu_{ij}^{\text{trans, naive}}$ underestimates $E[h^{-1}(y_{ij}^{*})|\mathbf{y}_s, \mathbf{X}_s]$. In order to get bias-corrected estimates, the best prediction on the original scale is needed.

In the case of a log-transformation, Berg and Chandra (2014) and Molina and Martín (2018) propose an analytical bias-correction. The best predictor for the out-of-sample units is defined for general transformations via an integral which can be solved analytically for $h() = log()$ by using $y_{ij}^{*}|\mathbf{y}_s, \mathbf{X}_s \sim \mathcal{N}\left(\mu_{ij}^{*}, \sigma_u^2(1-\gamma_i) + \sigma_e^2\right)$ - with corresponding density $f_{y_{ij}^{*}|\mathbf{y}_s,\mathbf{X}_s}$ - which comes directly from model (2.3),

$$\begin{aligned}\mu_{ij}^{\text{trans, bc}} = E[h^{-1}(y_{ij}^{*})|\mathbf{y}_s,\mathbf{X}_s] &= \int_{-\infty}^{+\infty} h^{-1}(x) f_{y_{ij}^{*}|\mathbf{y}_s,\mathbf{X}_s}(x)dx \\ &\overset{h^{-1}()=\exp()}{=} \exp\left(\mu_{ij}^{*} + \underbrace{\frac{\sigma_u^2(1-\gamma_i) + \sigma_e^2}{2}}_{=\alpha_i \text{ (bias-correction)}}\right).\end{aligned}$$

To the BLUP on the transformed scale (μ_{ij}^{*}) a bias-correction (α_i) is added before applying the back-transformation. $\mu_{ij}^{\text{trans, bc}}$ can be used to determine the bias-corrected estimator of the small area mean:

$$\hat{\overline{Y}}_i^{\text{trans, bc}} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} \hat{\mu}_{ij}^{\text{trans, bc}}\right) = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} \exp\left(\mathbf{x}_{ij}^T\hat{\beta} + \hat{u}_i + \hat{\alpha}_i\right)\right). \tag{2.4}$$

Molina and Martín (2018) propose for the MSE of $\hat{\overline{Y}}_i^{\text{trans, bc}}$ (2.4) both an analytical and a parametric bootstrap estimator. The package **saeTrafo** provides (2.4) and its bootstrap MSE estimator.

For $\hat{\overline{Y}}_i^{\text{trans, bc}}$ (2.4), out-of-sample population micro-data are needed which often causes problems with data confidentiality. Again, due to the Jensen's inequality a (second-order) bias is introduced if we use a naive back-transformation of the synthetic part (i.e., $\exp\left(\overline{\mathbf{x}}_i^T\hat{\beta}\right)$ instead of $\sum_{j\in\overline{s}_i}\exp\left(\mathbf{x}_{ij}^T\hat{\beta}\right)$). The estimator with first-order bias-correction (α_i) and naive back-transformation of the population-level aggregates is denoted by

$$\hat{\overline{Y}}_i^{\text{trans, bc-naive-agg}} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} \exp\left(\overline{\mathbf{x}}_i^T\hat{\beta} + \hat{u}_i + \hat{\alpha}_i\right)\right). \tag{2.5}$$

Due to the use of aggregated auxiliary data, this estimator has a second-order bias. To the best of my knowledge, no MSE estimator exists for $\hat{\overline{Y}}_i^{\text{trans, bc-naive-agg}}$ (2.5).

The next subsection presents small area means under the transformed NER model if only aggregated population-level auxiliary information is available. Therefore, it addresses the problem of limited data access and simultaneous transformation.

2.2.3 Small area means under limited auxiliary information

As emphasized in the previous subsection, the estimator $\hat{\overline{Y}}_i^{\text{trans, bc}}$ (2.4) requires population-level auxiliary data, which often leads to confidentiality constraints. In $\hat{\overline{Y}}_i^{\text{trans, bc-naive-agg}}$ (2.5), a second order bias remains because aggregated auxiliary data is used instead of individual data. In contrast to this, the method of Würz et al. (2022) aims to reduce the second-order bias due to the back-transformation of the synthetic part. Therefore, it offers a solution to deal with bias under limited auxiliary information while using log or log-shift transformation. This method approximates $\mathbf{x}_{ij}^T\hat{\beta}$ in the absence of population micro-data to reduce the second-order bias and combines this with the first-order bias-correction (α_i) for small area means.

Kernel density estimation for the synthetic part Due to limited auxiliary information, it is not possible to obtain $\left(\sum_{j\in\overline{s}_i}\exp\left(\mathbf{x}_{ij}^T\hat{\beta}\right)\right)$ necessary for computing $\hat{\overline{Y}}_i^{\text{trans, bc}}$ (2.4). Würz et al. (2022) propose an estimation method for the unknown synthetic part $(\mathbf{x}_{ij}^T\hat{\beta})$ under limited auxiliary information. They employ a KDE approach to estimate the distribution of $\mathbf{x}_{ij}^T\hat{\beta}$. This approach has two main advantages: firstly, the method of Würz et al. (2022) uses univariate KDE for the synthetic part $\left(\mathbf{x}_{ij}^T\hat{\beta}\right)$ instead of multivariate KDE to estimate the joint multivariate distribution of the auxiliary variables. Since current implementations of multivariate KDEs in R are restricted to a maximum number of auxiliary variables (cf. the widely used package **ks** (Duong, 2022) only allows for up to 6 covariates), many applications especially those with categorical data very quickly reach this limit. In contrasts, univariate KDE for the synthetic part avoids this restriction. Simulation studies in Würz et al. (2022) show that the estimation of the synthetic part is sufficient to reduce the second-order bias. Secondly, this method does not impose any parametric assumptions on the covariates and only require aggregated population-level auxiliary information.

KDE was first mentioned by Rosenblatt (1956) and Parzen (1962). Formally, KDE estimates the density f of a sample $X = \{X_1, ..., X_n\}$ by

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right), \tag{2.6}$$

where the function $k()$ is the kernel and h is the bandwidth. For more details on KDE, see for example Scott (2015). **saeTrafo** employs the Epanechnikov kernel (Epanechnikov, 1969), which is implemented using the `density` function of the **stats** package. Moreover, **saeTrafo** uses the method from Sheather and Jones (1991) for bandwidth selection.

As a first step, **saeTrafo** standardizes the predictions of the synthetic part from the NER model. For area i and individual j, the standardized predicted values z_{ij} are computed by

$$z_{ij} = \frac{\mathbf{x}_{ij}^T\hat{\beta} - \frac{1}{n_i}\sum_{j\in s_i}\mathbf{x}_{ij}^T\hat{\beta}}{\sqrt{\frac{1}{n_i}\sum_{j\in s_i}\left(\mathbf{x}_{ij}^T\hat{\beta} - \frac{1}{n_i}\sum_{j\in s_i}\mathbf{x}_{ij}^T\hat{\beta}\right)^2}}.$$

This formula employs the mean and the standard deviation from the sample data predictions of the synthetic part.

Second, the package adjusts the predictions with the help of aggregated population-level auxiliary data. It uses the mean $\overline{\mathbf{x}}_i^T\hat{\beta}$ and the empirical variation $\sigma_{i,\mathbf{X}^T\hat{\beta}} = \sqrt{\sum_{k=0}^{p}\sum_{l=0}^{p}\hat{\beta}_k\hat{\beta}_l \text{Cov}[\mathbf{x}_{ik},\mathbf{x}_{il}]}$, where $\text{Cov}[\mathbf{x}_{ik},\mathbf{x}_{il}]$ is the known covariance between the k-th and l-th explanatory variable for area i. This step incorporates the aggregated information from the census, which adds the SAE component to this method. Typically, in small area applications, sample sizes differ between areas. The package distinguishes between large sample sizes - standardized data (z_{ij}) from the respective area i (conditional) is used - and small sample sizes - standardized data (z_{ij}) from all areas (unconditional) is employed. In order to distinguish between large and small sample sizes, a threshold t is defined: for small sample sizes, i.e. below the threshold ($n_i < t$) - or even for an out-of-sample area - we use the standardized data from all areas to generate adjusted data for area i. The input values for the KDE (r_{im}) arise from the standardized values z_m. The index m ranges from $1, ..., n$ for sample sizes below t (unconditional) and from $1, ..., n_i$ for sample sizes above t (conditional). With

$$r_{im} = z_m\,\sigma_{i,\mathbf{X}^T\hat{\beta}} + \overline{\mathbf{x}}_i^T\hat{\beta} \qquad \text{for} \quad \begin{cases} m \in s & n_i < t \\ m \in s_i & n_i \geq t \end{cases} \tag{2.7}$$

we estimate the respective density using the KDE (2.6) for each area i. $\hat{f}_{h,i}$ denotes the resulting density for area i.

Small area means under limited auxiliary information In order to account for both types of biases the proposed method relies on the approximated area-specific density $\hat{f}_{h,i}$ of the synthetic part and the first-order bias-correction α_i:

$$\hat{\overline{Y}}_i^{\text{trans, bc}} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i}\exp\left(\hat{\mu}_{ij}+\hat{\alpha}_i\right)\right) \approx \frac{1}{N_i}\left(\underbrace{\sum_{j=1}^{N_i}\exp\left(\mathbf{x}_{ij}^T\hat{\beta}\right)}_{T_i}\exp\left(\hat{u}_i+\hat{\alpha}_i\right)\right).$$

$\hat{\mu}_{ij} = \mathbf{x}_{ij}^T\hat{\beta} + \hat{u}_i$ is defined as in the NER model. As shown above, under limited auxiliary information, the problem is reduced to determining the unknown back-transformed total (T_i). Würz et al. (2022) use numerical integration and the estimated density of the synthetic part $\hat{f}_{h,i}$ to determine the total $\hat{T}_i = \sum_{j=1}^{N_i}\exp\left(\mathbf{x}_{ij}^T\hat{\beta}\right) = N_i E[\exp(\mathbf{x}_{ij}^T\hat{\beta})] = N_i\int_{-\infty}^{+\infty}\exp(x)\hat{f}_{h,i}(x)dx$ from sample data and population-level auxiliary information - without using population micro-data. To achieve this, **saeTrafo** uses the package **sfsmisc** (Maechler et al., 2021). The requested small area estimator of the mean is obtained by inserting the estimated back-transformed area-specific totals $\hat{T}_i$:

$$\hat{\overline{Y}}_i^{\text{trans, bc-agg}} = \frac{1}{N_i}\hat{T}_i\exp\left(\hat{u}_i+\hat{\alpha}_i\right). \tag{2.8}$$

For the log-shift transformation, the characteristic shift-parameter $\hat{\lambda}$ is added

$$\hat{\overline{Y}}_i^{\text{trans, bc-agg}} = \frac{1}{N_i}\hat{T}_i\exp\left(\hat{u}_i+\hat{\alpha}_i\right) - \hat{\lambda}.$$

The R package **saeTrafo** is the first package providing these estimators to the users.

Uncertainty estimation For the estimator $\hat{\overline{Y}}_i^{\text{trans, bc-agg}}$ (2.8) under limited auxiliary data, Würz et al. (2022) develop a parametric bootstrap MSE that captures the additional uncertainty due to KDE and the estimation of the adaptive shift parameter in the case of a log-shift transformation. The following enumeration outlines the bootstrap procedure employed in **saeTrafo** for the log and log-shift transformation (these transformations are denoted with h).

1. Transform the data: $y_{ij}^* = h(y_{ij})$
2. Estimate $\hat{\beta}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ from sample data using model (2.3). In the case of the log-shift transformation, estimate $\hat{\lambda}$ as proposed by Rojas-Perilla et al. (2020).
3. For $b = 1, ..., B$
   (a) Generate $u_i^{(b)} \sim \mathcal{N}(0, \hat{\sigma}_u^2)$ and $e_{ij}^{(b)} \sim \mathcal{N}(0, \hat{\sigma}_e^2)$ for all areas i and $j \in s_i$.
   (b) Build bootstrap samples on the transformed scale

$$y_{ij}^{*(b)} = \mathbf{x}_{ij}^T \hat{\beta} + u_i^{(b)} + e_{ij}^{(b)}, \qquad \text{with } j \in s_i$$

   for all areas i and therefore determine the estimator $\hat{\overline{Y}}_i^{\text{trans, bc-agg, (b)}}$ (2.8) for all areas within each bootstrap replication b. Note, that λ is re-estimated within every replication b in case of the log-shift transformation.
   (c) Determine the true mean for each area i in each bootstrap replication b. Due to the lack of population micro-data for $\mathbf{x}$, an approximation of the true bootstrap mean is needed. From the available aggregated population-level values, Würz et al. (2022) construct an area-specific distribution on the transformed scale for each bootstrap replication b:

$$y_{ij}^{*(b)} | \mathbf{y}_s^{(b)}, \mathbf{X}_s, u_i^{(b)} \sim \mathcal{N}\left(\overline{\mathbf{x}}_i^T \hat{\beta} + u_i^{(b)}, \sigma_{i,\mathbf{X}^T\hat{\beta}}^2 + \hat{\sigma}_e^2\right), \tag{2.9}$$

   determine $\sigma_{i,\mathbf{X}^T\hat{\beta}} = \sqrt{\sum_{k=1}^p \sum_{l=1}^p \hat{\beta}_k \hat{\beta}_l \text{Cov}[\mathbf{x}_{ik}, \mathbf{x}_{il}]}$ from known covariances and estimated regression coefficients, and take $\hat{\sigma}_e^2$ from step 2. To get the true mean ($\overline{Y}_i^{(b)}$) on the original scale, Würz et al. (2022) combine the distributional assumptions on the transformed scale (2.9) with the properties of the exponential back-transformation function $h^{-1}() = exp()$, respectively $h^{-1}() = exp() - \lambda$:

$$\begin{aligned}
\overline{Y}_i^{(b)} &= \frac{1}{N_i} \sum_{j \in U_i} h^{-1}\left(y_{ij}^{*(b)}\right) | \mathbf{y}_s^{(b)}, \mathbf{X}_s, u_i^{(b)} \\
&\overset{h^{-1}()=\exp()}{=} \frac{1}{N_i} \sum_{j \in U_i} \exp\left(y_{ij}^{*(b)}\right) | \mathbf{y}_s^{(b)}, \mathbf{X}_s, u_i^{(b)} \\
&= \exp\left(\overline{\mathbf{x}}_i^T \hat{\beta} + u_i^{(b)} + 0.5\left(\sigma_{i,\mathbf{X}^T\hat{\beta}}^2 + \hat{\sigma}_e^2\right)\right).
\end{aligned}$$

For data-driven log-shift transformation, the analogue is

$$\overline{Y}_i^{(b)} = \exp\left(\overline{\mathbf{x}}_i^T\hat{\beta} + u_i^{(b)} + 0.5\left(\sigma^2_{i,\mathbf{X}^T\hat{\beta}} + \hat{\sigma}_e^2\right)\right) - \hat{\lambda},$$

where $\hat{\lambda}$ is the shift-parameter estimated from step 2.

4. Determine the MSE over the B bootstrap replications:

$$\widehat{\mathrm{MSE}}_i = \frac{1}{B}\sum_{b=1}^{B}\left(\overline{\hat{Y}}_i^{\text{trans, bc-agg, }(b)} - \overline{Y}_i^{(b)}\right)^2.$$

saeTrafo offers this parametric bootstrap procedure. To increase user-friendliness, it is possible to run this MSE estimation procedure on several cores. The expected execution times are displayed to the users.

The next section describes the Austrian data while Section 2.4 presents the core function `NER_Trafo`. The function provides the theory from this section in a user-friendly way, and demonstrates it based on the Austrian data.

2.3 Data sets for illustration

The main idea of SAE is to combine survey and population (census or administrative) data to increase the accuracy of the estimated indicator of interest. Since the target variable is only provided in the survey data, additional information from the population is used to support the prediction of the target variable using linear mixed models (Rao and Molina, 2015; Tzavidis et al., 2018). The package **saeTrafo** contains sample and population data to provide the users with exemplary data. The sample (`eusilcA_smp`) and population data (`eusilcA_pop`) are obtained from the package **emdi** (Kreutzmann et al., 2019). The authors provide an extensive description of the data generating processed of the `eusilcP` dataset coming from the package **simFrame** (Alfons et al., 2010). This household-level data set consists of synthetic Austrian European Union Statistics on Income and Living Conditions (EU-SILC) from 2006. For the package **emdi**, a spatially finer regional disaggregation was generated manually using a random assignment taking into account the different regional income-levels in Austria. The lowest regional level in this synthetic data set are the 94 Austrian districts. This population data comprises 25000 households, while there were more than 3.5 million households in Austria in 2006. The sample data is constructed by stratified random sampling and consists of 1945 households. The sample data includes 70 districts, leaving 24 areas out-of-sample. The equivalized household income (`eqIncome`) is the target variable and is only available within the sample. This variable is defined as the ratio of the total household disposable income and the equivalized household size. It was determined by the Organisation for Economic Co-operation and Development (OECD) (Hagenaars et al., 1994). In the following examples, 14 covariates serve as auxiliary data: `gender`, `eqsize`, `cash`, `self_empl`, `unempl_ben`, `age_ben`, `surv_ben`, `sick_ben`, `dis_ben`, `rent`, `fam_allow`, `house_allow`, `cap_inv`, and `tax_adj`. For detailed information, please refer to Kreutzmann et al. (2019). All 14 covariates are included within the sample and the full and aggregated population data. Furthermore,

the variable `district` is available in the data and represents the spatial target level.

The core function `NER_Trafo` of the package **saeTrafo** deals with different population data inputs. Figure 2.1 visualizes, which functions from the theory part (Section 2.2) applies under which population data input. To provide aggregates in a directly and user-friendly manner, `pop_area_size`, `pop_mean`, and `pop_cov` are available as data sources in the package. All three data objects are calculated from `eusilcA_pop`. Their direct availability makes it more convenient for the user to try out all functionalities of **saeTrafo**.

2.4 Core functionalities

This section is structured accordingly: Section 2.4.1 gives an overview of the main function `NER_Trafo`, Section 2.4.2 shows how `NER_Trafo` is applied using the exemplary data, and Section 2.4.3 demonstrates the possibilities of **saeTrafo**'s generic functions to analyse, visualize, and export the corresponding S3 object.

2.4.1 Overview NER_Trafo

The `NER_Trafo` function provides the methodology from Section 2.2. `NER_Trafo` has 16 input arguments, takes the different data input possibilities into account, and allows for a variety of specifications (cf. Table 2.1). As a minimum input, the sample data (`smp_data` and `smp_domains`), the formula object (`fixed`), and population data - either the aggregated data (`pop_area_size`, `pop_mean`, and optional `pop_cov`) or the individual data (`pop_data` and `pop_domains`) - must be entered. As **saeTrafo** uses the S3 object system, `NER_Trafo` returns an object of class `saeTrafo` and `NER` (Chambers and Hastie, 1992). The reason for assigning two classes to the object is ability to integrate further SAE models in future releases. The output object consists of ten components. In this way, the user can directly access the point estimates (`ind`), the uncertainty estimates (`MSE`), transformation parameters (`transform_param`), information on the underlying linear mixed-effects model as in the package **nlme** (Pinheiro et al., 2022) (`model`), a list describing the data input (`framework`), the selected transformation (`transformation`), the method for transformation parameter estimation (`method`), the formula object (`fixed`), the function call (`call`), and number of successful bootstraps for bootstrap MSE estimation procedures (`successful_bootstraps`).

Figure 2.1 illustrates which estimation methods for point and MSE estimation are used under different combinations of selected transformation and type of population data. If no transformation is selected, **saeTrafo** employs the classical model by Battese et al. (1988). Since no individual data are necessary, potentially used population micro-data are processed into aggregates in a first step. Under the log or log-shift transformation **saeTrafo** automatically selects between different methods depending on the data. **saeTrafo** uses the estimator of Würz et al. (2022) if population aggregates (means, covariances, and populations area sizes) are supplied in the presence of transformations. If only means and area sizes under log or log-shift transformation are present, the `NER_Trafo` function employs the estimator $\hat{\overline{Y}}_i^{\text{trans,bc-naive-agg}}$ (2.5) for which no MSE estimator exists. This estimator only corrects the first-order bias and neglects

Table 2.1: Input arguments of function `NER_Trafo`.

Arguments	Short description	Default
`fixed`	Formula object with fixed effects and response variable of the NER model	
`pop_area_size`	Population sizes per domain	NULL
`pop_mean`	Population means for all fixed effects per domain	NULL
`pop_cov`	Population covariance matrices between all fixed effects per domain	NULL
`pop_data`	Census or administrative data containing all fixed effects	NULL
`pop_domains`	Domain identifier for population data	NULL
`smp_data`	Survey data comprising the fixed effects and the response variable	
`smp_domains`	Domain identifier for sample data	
`threshold`	Threshold for using pooled domain data	30
`B`	Number of bootstrap replications for bootstrap MSE estimation	50
`transformation`	Type of transformation: no, log, log-shift	log-shift
`interval`	Interval for estimating the optimal parameter of log-shift transformation	range of response
`MSE`	MSE estimation	FALSE
`parallel_mode`	Mode of parallelization for bootstrap MSE procedure	automatic
`cpus`	Kernels for parallelization for bootstrap MSE procedure	1
`seed`	Seed for random number generator within bootstrap MSE procedure	123

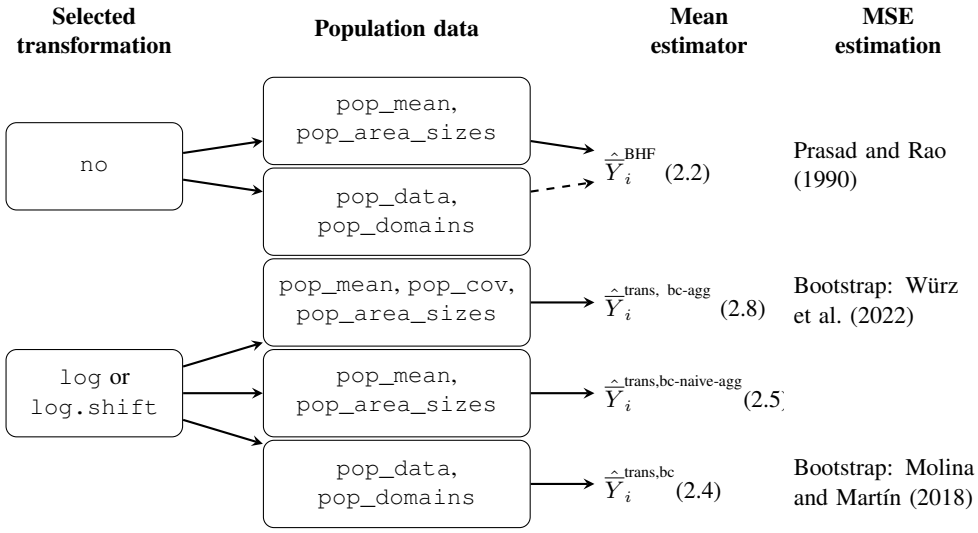


Figure 2.1: Overview of different estimation methods provided in function `NER_Trafo`. These estimation methods are chosen depending on the selected transformation and the type of provided population data.

the second bias due to limited data. An alternative method - not implemented in R yet - is the estimator from Li et al. (2019), for which no MSE estimator exists too. If the log or log-shift transformation occur with individual population data, **saeTrafo** uses the estimator $\hat{\overline{Y}}_i^{\text{trans,bc}}$ (2.4) together with its bootstrap MSE. Please note, that in the cases of individual population data other packages like **emdi** (Kreutzmann et al., 2019) provide further functionalities: the estimation of quantiles, inequality indicators, and further transformations (box-cox transformation (Box and Cox, 1964) and dual transformation (Yang, 2006)). These options become available in the `ebp` function of **emdi** which applies the method of Molina and Rao (2010). Since the `ebp` function is based on Monte Carlo replications, the run time is longer than for `NER_Trafo`.

2.4.2 Estimation of (transformed) nested error regression models

Synthetic Austrian EUSILC data (cf. Section 2.3) is used to illustrate the functionalities of **saeTrafo** and the estimation with `NER_Trafo`. The example demonstrates the estimation of the small area means for the equivalized household income (`eqIncome`) at the disaggregation level of 94 Austrian districts. The sample, population, and aggregated data are available in **saeTrafo**:

```
R> library(saeTrafo)
R> data("eusilcA_pop")
R> data("eusilcA_smp")
R> data("pop_area_size")
R> data("pop_mean")
R> data("pop_cov")
```

The data allow for easy testing of the different methods implemented and bundled in `NER_Trafo`. For illustration purposes, the example focuses on estimating $\hat{\overline{Y}}_i^{\text{trans, bc-agg}}$ (2.8), therefore it is sufficient to insert only aggregated population data. In addition to the point estimates, MSE estimates are calculated too, so `MSE` is set to `TRUE`. Furthermore, the setting for the `threshold` for pooled estimation (cf. (2.7)) is set to 50. To prevent long run times for MSE estimation the default of the number of bootstrap replications is only 50, whereby parallelization is available in the function. To obtain a more precise MSE estimate, `B` is increased to 250 in the example. The `seed` is set to 2022 to ensure reproducibility of the results.

```
R> formula <- eqIncome ~ gender + eqsize + cash + self_empl +
+    unempl_ben + age_ben + surv_ben + sick_ben + dis_ben +
+    rent + fam_allow + house_allow + cap_inv + tax_adj

R> NER_model <- NER_Trafo(fixed = formula,
+    pop_area_size = pop_area_size, pop_mean = pop_mean,
+    pop_cov = pop_cov, smp_data = eusilcA_smp,
+    smp_domains = "district", B = 250, threshold = 50,
+    MSE = TRUE, seed = 2022)
```

The R object `NER_model` is of two classes `saeTrafo` and `NER`. For this S3 object several generic functions are provided within **saeTrafo** and presented in the following section.

2.4.3 Generic functions

The most important generic functions of the R package **saeTrafo** (summary output, diagnostic plots, visualisation of estimates, and their export) are shown in detail. All other functionalities are only briefly introduced.

Summary of a saeTrafo object By applying the `summary` function on an object of class `saeTrafo`, R-user receive basic information and first diagnostic results. In addition to the call, small area specific characteristics (number of out-of-sample and in-sample domains, information on sample sizes, and their distribution among domains) are displayed. To assess the proportion of variance explained by the model, **saeTrafo** provides both a marginal and conditional R^2 following Nakagawa and Schielzeth (2013). The R^2s are implemented as in the **emdi**-package (Kreutzmann et al., 2019) and use the **MuMIn**-package from Barton (2018). Moreover, the output shows information on the residual diagnostics for the unit-level errors (e_{ij}) and the domain-specific random effects (u_i). If a transformation is selected, **saeTrafo** calculates these diagnostics on the transformed scale and hence help to judge, if the transformation assists to meet the normality assumption of both components. The ICC relates the variances (σ_u^2 and σ_e^2) to each other. Finally, the `summary` function outputs information on the transformation and the selected parameter λ.

```
R> summary(NER_model)

Nested Error Regression Model

Call:
 NER_Trafo(fixed = eqIncome ~ gender + eqsize + cash +
    self_empl + unempl_ben + age_ben + surv_ben + sick_ben +
    dis_ben + rent + fam_allow + house_allow + cap_inv +
    tax_adj,
    pop_area_size = pop_area_size, pop_mean = pop_mean,
    pop_cov = pop_cov, smp_data = eusilcA_smp,
    smp_domains = "district", threshold = 50, B = 250,
    MSE = TRUE, seed = 2022)

Out-of-sample domains:  24
In-sample domains:  70

Sample sizes:
Units in sample:  1945
Units in population:  25000
```

```
                    Min. 1st Qu. Median      Mean 3rd Qu. Max.
Sample_domains        14    17.0   22.5  27.78571   29.00  200
Population_domains     5   126.5  181.5 265.95745  265.75 5857

Explanatory measures:
 Marginal_R2 Conditional_R2
   0.6233538      0.7054886

Residual diagnostics:
               Skewness Kurtosis Shapiro_W    Shapiro_p
Error         0.6222910 7.607189 0.9706711 1.705890e-19
Random_effect 0.4788713 2.726898 0.9737695 1.487627e-01

ICC:  0.2180689

Transformation:
 Transformation Method Optimal_lambda
      log.shift   reml       27907.57
```

The output of the example shows that the synthetic Austrian data consists of 24 out-of-sample domains and 70 in-sample domains. As the sample sizes over domains are considerably small (Median: 22.5) this is a classical small area problem. Both the marginal and conditional coefficients of determination are high with values above 62%. The normality assumption for the random effects is not rejected at a significance level of 5%. For the individual errors, this assumption is rejected with p = 1.705890e-19. The random effects contribute to around 21% of the model variance. The chosen transformation is the log-shift transformation with REML-estimated transformation parameter of $\lambda = 27907.57$.

Diagnostic plots for the nested error regression model The `plot` function provides five plots bundling the most important diagnostic information: Q-Q plots to judge the normality assumption on the error terms (cf. Figure 2.2a), the deviation of both the density from the normal distribution for the individual errors (cf. Figure 2.2b) and the random effects (cf. Figure 2.2c), the Cook's distance to identify outliers (cf. Figure 2.2d) as well as information on the optimal transformation parameter λ for the log-shift transformation (cf. Figure 2.2e). The `plot` function allows customized settings: the input arguments `label`, `color`, `cooks`, and `range` enable direct changes to the plots. In addition, with `gg_theme` there is the possibility of further personalisation of the plots by using the **ggplot2** package (Wickham, 2016).

```
R> plot(NER_model)
```

In the Austrian income example, the Q-Q plot (cf. Figure 2.2a) and the density plot (cf. Figure 2.2c) confirm the normality assumptions of the underlying model for the random effects. However, for the individual error term, the Q-Q plot (cf. Figure 2.2a) shows several outliers.

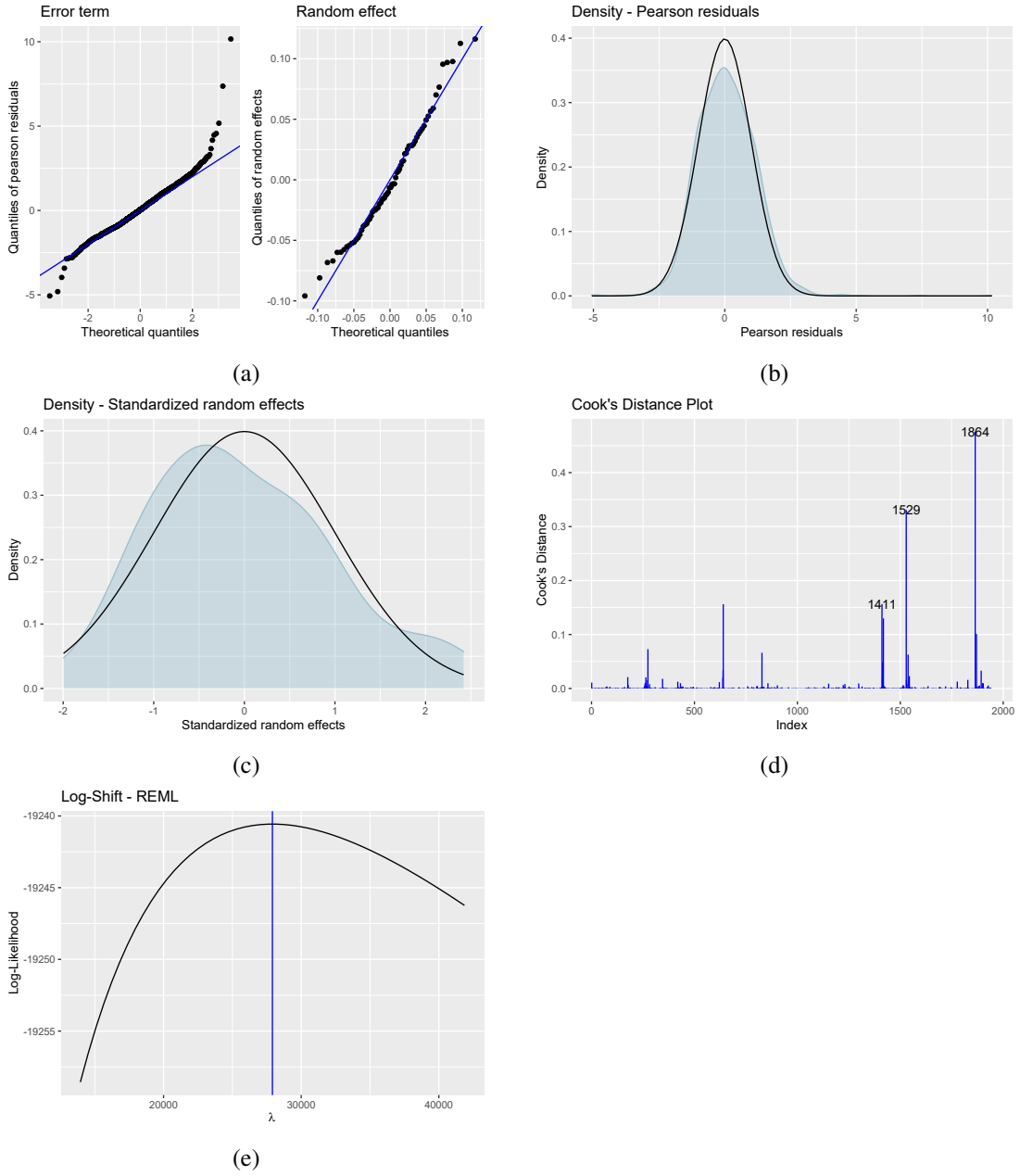


Figure 2.2: Diagnostic plots from generic function `plot`: Q-Q plots (a) and two density plots ((b) and (c)) to check the normality assumption for both error terms, Cook's distance plot for detecting potential outliers (d), and log-likelihood for the optimal shift parameter λ (e).

The Cooks distance plot highlights three individuals as possible outliers. The last plot (cf. Figure 2.2e) shows the log-likelihood reaching its maximum at $\lambda = 27907.57$. This plot is only supplied for the log-shift transformation.

Comparing point and optional MSE/CV estimates The generic function `compare_plot` is very important for users to evaluate the quality of their model-based estimates. In SAE applications the comparison of the particular model-based estimator to the respective direct estimator is of central importance. Since **saeTrafo** does not provide a function for determining direct estimators, other packages must be utilized. Among others the **survey** package (Lumley, 2004), the **laeken** package (Alfons and Templ, 2013), and the **emdi** package (Kreutzmann et al., 2019) enable the estimation of disaggregated direct estimators and their variances from a survey. Up to now, the generic function `compare_plot` works only with direct estimators from the package **emdi**. The procedure for this is shown in the exemplary code. For the comparison of point estimates, `compare_plot` returns two types of plots: a scatter plot following Brown et al. (2001) and a lineplot with direct and model-based domain-wise estimates. To compare the uncertainty - if `MSE` or `CV` is set to `TRUE` - `compare_plot` returns a boxplot and a scatterplot. In addition to a direct adjustment of the visualisation with `label`, `color`, `shape`, and `line_type` the argument `gg_theme` offers the possibility for further visualisation options using the **ggplot2** package (Wickham, 2016).

```
R> require(emdi)
R> library(emdi)
R> emdi_direct <- direct(y = "eqIncome",
+    smp_data = eusilcA_smp, smp_domains = "district",
+    weights = "weight", var = TRUE,
+    na.rm = TRUE)
R> detach("package:emdi", unload = TRUE)

R> compare_plot(model = NER_model, direct = emdi_direct,
+    CV = TRUE)
```

Both plots comparing direct and model-based point estimates show that the direct and model-based estimates are close to each other, as the regression line and the identity line are close to each other (cf. Figure 2.3a) and the model-based estimates track the direct ones (cf. Figure 2.3b). Furthermore, the CV is assessed in Figure 2.3c and 2.3d. As the boxplots show, the uncertainty - measured by the CV - is reduced clearly. The scatterplot which orders the domains by their sample size (from low to high) supports this impression.

Visualization of regional disaggregated estimates on a map The spatial visualisation on a map is simplified considerably by the `map_plot` function which generates maps automatically if a `SpatialPolygonsDataFrame` from package **sp** (Bivand et al., 2013) is provided additionally to the S3 object from `NER_Trafo`. As in **emdi** (Kreutzmann et al., 2019), the same polygon data showing Austrian districts is available within **saeTrafo**, so that it is possible to visualize the estimates on a map. The `load_shapeaustria` function loads this map and

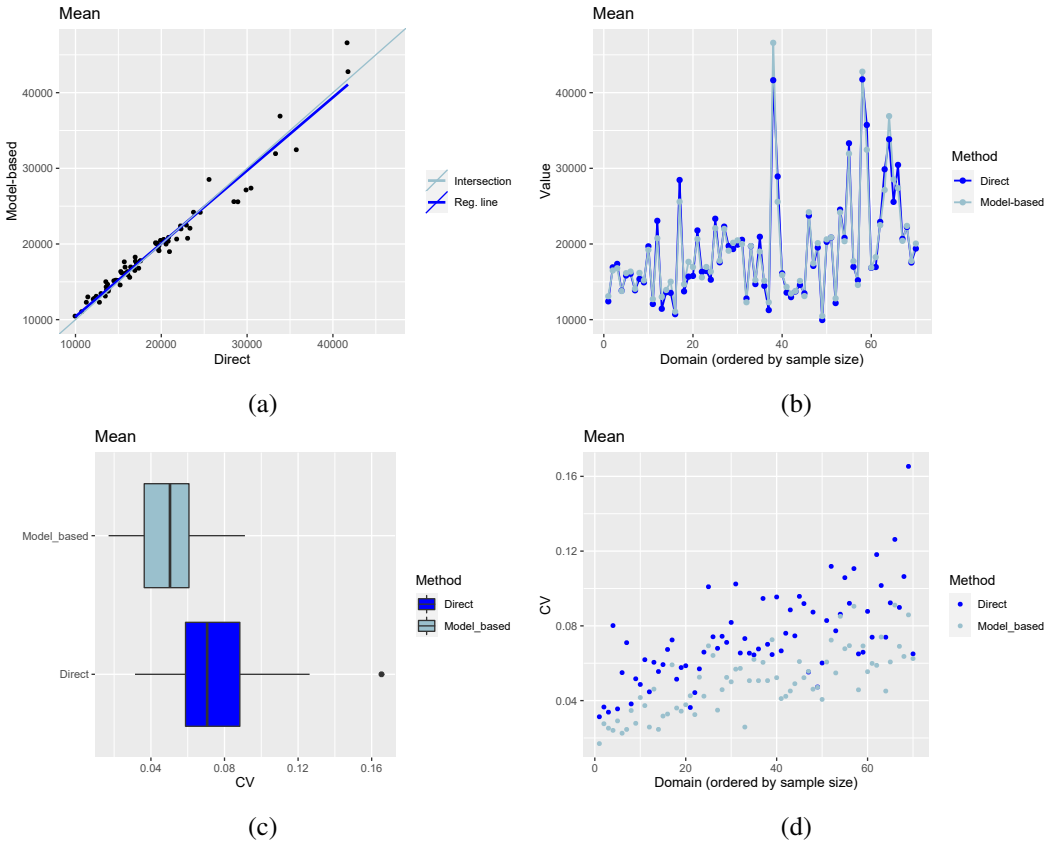


Figure 2.3: Plots for comparison to direct estimates from generic function `compare_plot` for the NER model: scatter plot (a), line plots with estimates ordered by domain-specific sample size (b), boxplots to compare CV for both estimators (c), and scatter plot for CV estimates ordered by domain-specific sample size (d).

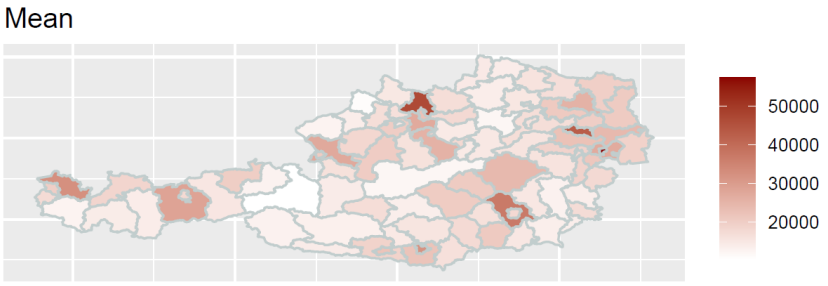


Figure 2.4: Map with Austrian districts showing their small area means for the equivalized household income from function `map_plot`.

the `map_plot` function offers various options for the users. This function directly supplies settings for the graphical representation (`color`, `scale_points`, and `guide`), outputs the processed data (`return_data`), and enables options to customize the map with the help of **ggplot2** (Wickham, 2016). If the domain IDs within the `SpatialPolygonsDataFrame` and the S3 object differ, `map_tab` enables the entry of a `data.frame` for the assignment of the domain IDs.

```
R> load_shapeaustria()

R> map_plot(NER_model, map_obj = shape_austria_dis,
+         map_dom_id = "PB")
```

The map in Figure 2.4 shows the mean equivalized household income for all 94 Austrian districts produced by the SAE methods explained above. Smaller values are mostly in rural districts (like Zell am See with the lowest value of 10469.93€) and higher mean equivalized household incomes appear in more urban districts.

Exporting the results and most important model information In addition to the evaluation and visualization of the point estimates (and uncertainty estimates), the package enables the export to other software. **saeTrafo** offers direct and user-friendly export of the estimates and the information from the `summary` function on the **saeTrafo** object to the software Excel.

```
R> write.excel(NER_model, file = "excel_output.xlsx",
+         CV = TRUE)
```

In addition, the export to `OpenDocument` format is also supported.

```
R> write.ods(NER_model, file = "excel_output.xlsx", CV = TRUE)
```

In both functions it can be specified if the CVs and MSEs should also be exported. If `split` is set to `TRUE`, the point estimators, MSEs and CVs are saved in separate worksheets, respectively separate documents. The created files are stored in the working directory.

Further generic functions Besides the generic functions already presented in detail, **saeTrafo** offers further generics: the function `estimators` is convenient to get point, MSE and CV

estimates. In addition, the widely known functions `as.data.frame`, `as.matrix`, `head`, `print`, `subset`, and `tail` can be applied to the S3 object created with `estimators`. The `print` function returns the most important model information. To facilitate the comparison between SAE estimators, the generic function `compare_pred` exists and creates a data set with point or MSE estimators of both objects. To also enable comparisons with other SAE methodology, an **emdi** object can be entered.

To further increase user-friendliness, well-known, and widely used generic functions from the **stats** package can be used with **saeTrafo**. Thus, the following functions can be applied to the S3 object of `NER_Trafo`: `coef`, `confint`, `family`, `fitted`, `formula`, `logLik`, `nobs`, `predict`, `residuals`, `sigma`, `terms`, and `vcov`.

Since the linear mixed models used are calculated with the **nlme** package (Pinheiro et al., 2022), the following generic functions for **nlme** objects are available for the S3 object of **saeTrafo**: `fixef`, `getData`, `getGroups`, `getGroupsFormula`, `getResponse`, `getVarCov`, `intervals`, and `ranef`.

2.5 Conclusion

The main focus of **saeTrafo** is to make the new methodology by Würz et al. (2022) publicly available. This methodology resolves the problem of not having access to individual population data while using transformations in the context of unit-level small area models. This method and its uncertainty estimation are supplied by the function `NER_Trafo`. In addition, the package provides the following methods: the well-known estimator by Battese et al. (1988), the bias-corrected estimator from Molina and Martín (2018) using population micro-data, and a first-order bias-corrected estimator using aggregated population data. An advantage of this function is the appropriate and automatic selection of small area methodology under different possible data inputs and transformations (none, log, and data-driven log-shift transformation). **saeTrafo** guarantees user-friendliness by providing all methods and their respective MSE (including parallelization options) within the `NER_Trafo` function. For this S3 object, a variety of generic functions are offered. They automate the creation of important plots for model diagnostics and the assessment of the estimator's quality. Furthermore, options for visualizing the estimates on maps and the export of estimators are provided. Further generic functionalities increase the user-friendliness.

This last paragraph outlines possible new features of **saeTrafo** for future releases: The choice between different methodologies to estimate the MSE will increase user-friendliness. For the estimator $\hat{\overline{Y}}_i^{\text{trans,bc}}$, Molina and Martín (2018) propose an analytical MSE in addition to the bootstrap version already supplied in **saeTrafo**. Further releases would profit by including this version. Moreover, **saeTrafo** offers the MSE of Prasad and Rao (1990) for the classical NER model. Further MSE estimating options are desirable. To have a MSE for the first-order bias-corrected estimator (trans, bc-naive-agg), theoretical research is first necessary. Including alternative SAE methods such as the method of Li et al. (2019) will increase the flexibility of the package. Overall, the **saeTrafo** software package is written in such a way that this can be easily extended with other small area model classes. For long-term future versions, this is

aspired.

Acknowledgements

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

Chapter 3

Analysing opportunity cost of care work using mixed effects random forests under aggregated census data

3.1 Introduction

Evidence-based policy requires reliable empirical information on social and economic conditions summarised by appropriate indicators. For questions addressing regional and spatial aspects of inequality, we need precise and reliable information extending beyond aggregate levels into highly disaggregated geographical and other domains (e.g., demographic groups). An apparent trade-off regarding the work with survey data is the inverse relation between high spatial resolution and decreasing sample sizes on the level of interest. The estimation of indicators under these circumstances can be facilitated using an appropriate model-based methodology collectively referred to as Small Area Estimation (SAE) (Rao and Molina, 2015; Tzavidis et al., 2018).

Models handling unit-level survey data for the estimation of area-level means are predominantly regression-based linear mixed models (LMM), where the hierarchical structure of observations is captured by random effects. A well-known example is the nested error regression model (Battese et al., 1988) - further labelled as BHF - which requires access to the survey and to area-level auxiliary information. A versatile extension of the BHF model is the EBP approach by Molina and Rao (2010) with which even non-linear indicators can be estimated and, unlike the BHF, requires access to population-level auxiliary data. The underlying LMM of the BHF (and the EBP) relies on distributional and structural assumptions that are prone to violations in SAE applications. Working with social and economic inequality data in LMMs requires assumptions of linearity and normality of random effects and error terms, which hardly meet empirical evidence. Jiang and Rao (2020) remind, that optimality results and predictive performance of model-based SAE are inevitably connected to the validity of model assumptions. Without theoretical and practical considerations regarding violated assumptions, estimates are potentially biased and mean squared error (MSE) estimates are unreliable.

In SAE, several strategies evolved to prevent model-misspecification: A well-known ex-

ample is the assurance of normality by transforming the dependent variable (Sugasawa and Kubokawa, 2017; Tzavidis et al., 2018; Sugasawa and Kubokawa, 2019; Rojas-Perilla et al., 2020). Furthermore, the use of models under more flexible distributional assumptions is a fruitful approach (Diallo and Rao, 2018; Graf et al., 2019). From a different perspective, semi- or non-parametric approaches for the estimation of area-level means are investigated among others by Opsomer et al. (2008), using penalized spline components within the LMM setting. A distinct methodological option to avoid the parametric assumptions of LMMs are machine learning methods. These methods are not limited to parametric models and learn predictive relations from data, including higher order interactions between covariates, without explicit model assumptions (Hastie et al., 2009; Varian, 2014). Recently, Krennmair and Schmid (2022) introduce a framework enabling a coherent use of tree-based machine learning methods in SAE. They propose a non-linear, data-driven, and semi-parametric alternative for the estimation of area-level means by using mixed effects random forests (MERF) in the methodological tradition of SAE. In general, random forests (RF) (Breiman, 2001) exhibit excellent predictive performance in the presence of outliers and implicitly solve problems of model-selection (Biau and Scornet, 2016). MERFs (Hajjem et al., 2014) combine these advantages with the ability to model hierarchical dependencies.

All previously mentioned model-based strategies against model-misspecification in SAE assume access to auxiliary information from population-level micro-data. Due to data security reasons, the access to unit-level census or register data is limited, which imposes a strong restriction for researchers and SAE practitioners. However, aggregated population-level auxiliary data (e.g., means) are often available at finer spatial resolution.

In this paper, we present a methodology for the estimation of area-level means using MERFs under limited population-level auxiliary information. We propose a purely data-driven approach for solving the dual problem (model-misspecification and limited auxiliary data). Particularly, we introduce a strategy for the adaptive incorporation of auxiliary information through calibration-weights for the estimation of area-level means. The determination of weights without explicit distributional assumptions is based on the empirical likelihood (EL) approach (Chen and Qin, 1993; Qin and Lawless, 1994; Han and Lawless, 2019). For the point estimation of area-level means, Li et al. (2019) propose the use of EL-based calibration weights and introduce a bias-corrected transformation approach using aggregated covariate data combined with the smearing approach of Duan (1983). Complementing our proposed method for point estimates, we introduce a non-parametric bootstrap estimator assessing the uncertainty of estimated area-level means. To the best of our knowledge, no comparable procedure exists for uncertainty estimation in the context of non-linear semi-parametric tree-based procedures under limited data access. We highlight strengths and weaknesses of our approach for point and uncertainty estimates by comparing it to existing SAE methods under limited auxiliary information in a model-based simulation.

We demonstrate our methodology using the 2011 Socio-Economic Panel (SOEP) (Socio-Economic Panel, 2019) combined with aggregate census information from the same year to estimate the average individual opportunity cost of care work for 96 regional planning regions (RPRs) in Germany. We refer to care work as unpaid working hours attributed to child- or

elderly-care reported by the SOEP. Opportunity cost is an economic concept comprising the time allocation problem, where the time allocated for care work implicitly corresponds to time not providing paid work (Buchanan, 1991). Informally provided care work has no direct corresponding monetary value and the determination of a correct shadow-price for the economic value is difficult. Classical interpretations of labour supply in economics such as Becker (1965) imply that an individual's hourly wage is an acceptable approximation to the unknown opportunity cost of time for working population. Thus, we measure time cost by multiplying an individual's care time by the opportunity cost of the person's time represented as the reported hourly wage calculated also from reported income in the SOEP data. We are aware that our application is at best a first approximation making regional differences in opportunity cost of care work visible, accountable, and comparable. Unpaid care work mitigates public and private expenses on needed health services and infrastructure (Charles and Sevak, 2005). On the other hand, care-giving has a complex impact on the labour market (Truskinovsky and Maestas, 2018; Stanfors et al., 2019), for instance by affecting workforce individuals through personal or social burdens (Bauer and Sousa-Poza, 2015). From a macro-perspective, several studies examine the economic value of care work for countries through the concept of opportunity cost (Chari et al., 2015; Ochalek et al., 2018; Mudrazija, 2019) and provide empirical evidence for policy measures.

While the mapping of spatial patterns of income inequality in Germany is of scientific interest (Frick and Goebel, 2008; Kosfeld et al., 2008; Fuchs-Schündeln et al., 2010), to the best of our knowledge, no study on regional dispersion of opportunity cost of unpaid care work exists. From a spatial perspective, Oliva-Moreno et al. (2019) provide estimates on the economic value of time of informal care for two regions in Spain. We maintain that mapping opportunity cost of care work in Germany is particularly interesting given the German history of Reunification and the German Federalism, characterized by powerful regional jurisdictions and different laws for aspects directly affecting care work. The visualization of opportunity cost highlights regional patterns, adding insights for planning and comparison of social-compensation policies.

The rest of the paper is structured as follows: Section 3.2.1 states a general mixed model that treats LMMs in SAE as special cases and enables the use of tree-based models. We consider the estimation of area-level means using MERFs, which effectively combine advantages of non-parametric random forests with the possibility to account for hierarchical dependencies. Section 3.2.2 describes our area-level mean estimator based on MERFs under limited data access. We scrutinize the use of EL calibration weights and subsequently address methodological limitations in Section 3.2.3. As a result, we propose a best practice strategy to ensure the proper usability of EL calibration weights in the context of SAE. Section 3.3 introduces a non-parametric bootstrap-scheme for the estimation of the area-level MSE. In Section 3.4, we use model-based simulations under complex settings to assess the performance of our stated methods for point and MSE estimates, showing that MERFs are a valid alternative to existing methods for the estimation of SAE means under limited data access. In Section 3.5, we estimate the average individual opportunity cost of care work for 96 RPRs in Germany using the 2011 SOEP data. After the introduction of data sources and direct estimates in Section 3.5.1, we highlight modelling and robustness properties of our proposed methods for point and

uncertainty estimates compared to direct and other SAE estimates under limited auxiliary data. In Section 3.6, we conclude and motivate further research.

3.2 Theory and method

This section introduces a general mixed model enabling a simultaneous discussion of traditional LMM-based models in SAE such as the model of Battese et al. (1988) as well as semi-parametric interpretations such as the model of Krennmair and Schmid (2022) using MERFs. Section 3.2.2 provides details on our proposed methodology for MERFs under limited covariate data access and the determination of area-specific calibration weights based on EL. We close the section with a discussion on limitations of EL for SAE and state a best practice strategy ensuring the usability of our proposed point estimator in challenging empirical examples.

3.2.1 Model and estimation of coefficients

We assume a finite population U of size N consisting of D separate domains $U_1, U_2, ..., U_D$ with $N_1, N_2, ..., N_D$ units, where index $i = 1, ..., D$ indicates respective areas. The continuous target variable y_{ij} for individual observation j in area i is available for every unit within the sample. Sample s is drawn from U and consists of n units partitioned into sample sizes $n_1, n_2, ..., n_D$ for all D areas. We denote by s_i the sub-sample from area i. The vector $\mathbf{x}_{ij} = (x_1, x_2, ..., x_p)^\intercal$ includes p explanatory variables and is available for every unit j within the sample s. The relationship between $\mathbf{x}_{ij}$ and y_{ij} is assumed to follow a general mixed effects regression model:

$$y_{ij} = f(\mathbf{x}_{ij}) + u_i + e_{ij} \quad \text{with} \quad u_i \sim N(0, \sigma_u^2) \quad \text{and} \quad e_{ij} \sim N(0, \sigma_e^2). \tag{3.1}$$

Function $f(\mathbf{x}_{ij})$ models the conditional mean of y_{ij} given $\mathbf{x}_{ij}$. The area-specific random effect u_i and the unit-level error e_{ij} are assumed to be independent. For instance, defining $f(\mathbf{x}_{ij}) = \mathbf{x}_{ij}^\intercal \beta$ with $\beta = (\beta_1, ..., \beta_p)^\intercal$ coincides with the well-known nested error regression model of Battese et al. (1988) labelled as BHF. An empirical best linear unbiased predictor for the area-level mean μ_i can be expressed as:

$$\hat{\mu}_i^{\text{BHF}} = \bar{\mathbf{x}}_i^\intercal \hat{\beta} + \hat{u}_i,$$

where $\bar{\mathbf{x}}_i = \frac{1}{N_i} \sum_{j \in U_i} \mathbf{x}_{ij}$ denotes area-specific population means on p covariates. In a variety of real-world examples, required assumptions for the BHF model hardly meet empirical evidence. Apart from transformation strategies to meet the required assumptions, non-parametric approaches can be used alternatively (Jiang and Rao, 2020). Tree-based machine learning methods such as RFs (Breiman, 2001) are data-driven procedures identifying predictive relations from data, including higher order interactions between covariates, without explicit model assumptions (Hastie et al., 2009; Varian, 2014). RFs inherently perform model-selection and properly handle the presence of outliers (Biau and Scornet, 2016). However, an implicit assumption of tree-based models is the required independence of unit-level observations.

Defining f in Model (3.1) to be a RF results in a semi-parametric framework, combining

advantages of RFs with the ability to model hierarchical structures of survey data using random effects. Krennmair and Schmid (2022) estimate area-level means with RFs (Breiman, 2001) introducing a method that enables the estimation of model-components $\hat{f}$, $\hat{u}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ in the context of SAE. The so-called mixed effects random forest (MERF) uses a procedure reminiscent of the EM-algorithm (Hajjem et al., 2014). For fitting Model (3.1) (where f is a RF) on survey data, the MERF algorithm subsequently estimates a) the forest function, assuming the random effects term to be correct and b) estimates the random effects part, assuming the Out-of-Bag-predictions (OOB-predictions) from the forest to be correct. OOB-predictions utilize the unused observations from the construction of each forest's sub-tree (Breiman, 2001; Biau and Scornet, 2016). The estimation of variance components $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_u^2$ is obtained implicitly by taking the expectation of ML estimators given the data. For further methodological details, we refer to Krennmair and Schmid (2022). The resulting estimator for the area-level mean for MERFs is summarized as:

$$\hat{\mu}_i^{\text{MERF}} = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \hat{u}_i = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i} \left(\frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - \hat{f}(\mathbf{x}_{ij})) \right), \tag{3.2}$$

$$\text{where} \quad \bar{\hat{f}}_i(\mathbf{x}_{ij}) = \frac{1}{N_i} \sum_{j \in U_i} \hat{f}(\mathbf{x}_{ij}).$$

3.2.2 MERFs under aggregated data

Estimates for the area-level mean μ_i using MERFs from Equation (3.2) require unit-level auxiliary census data as input for f. In contrast to the linear BHF model by Battese et al. (1988), aggregated covariate data cannot directly be used for non-linear or non-parametric procedures such as RFs, as in general $f(\bar{\mathbf{x}}_i) \neq \bar{f}_i(\mathbf{x}_{ij})$. Although the access to auxiliary population micro-data for the covariates imposes a limitation for practitioners, not many methods in SAE cope with the dual problem of providing robustness against model-failure, while simultaneously working under limited auxiliary data (Jiang and Rao, 2020). We propose a solution overcoming this issue by calibrating model-based estimates from MERFs in Equation (3.2) with weights that are based only on aggregated census-level covariates (means). The general idea originates from the bias-corrected transformed nested error regression estimator using aggregated covariate data (*TNER2*) by Li et al. (2019). We build on their idea of using calibration weights for SAE based on EL (Owen, 1990; Qin and Lawless, 1994; Owen, 2001) and transfer it to MERFs. As a result, our proposed method offers benefits of RFs such as robustness and implicit model-selection, while simultaneously working in cases of limited access to auxiliary covariate data. In short, our estimator for the area-level mean can be written as:

$$\hat{\mu}_i^{\text{MERFagg}} = \sum_{j=1}^{n_i} \hat{w}_{ij} \left[\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i \right]. \tag{3.3}$$

Note that optimal estimates for required model-components $\hat{f}$ and $\hat{u}_i$ are obtained similar to Equation (3.2) from survey data using the MERF algorithm as described by Krennmair and Schmid (2022). We incorporate aggregate census-level covariate information through the cali-

bration weights w_{ij}, which balance unit-level predictions to achieve consistency with the area-wise covariate means from census data. Following Owen (1990) and Qin and Lawless (1994) the technical conditions for w_{ij} are to maximize the profile EL function $\prod_{j=1}^{n_i} w_{ij}$ under the following three constraints:

- $\sum_{j=1}^{n_i} w_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i}) = 0$, monitoring the area-wise sum of distances between survey data and the population-level mean, denoted as $\bar{\mathbf{x}}_{\text{pop},i}$, for auxiliary covariates;
- $w_{ij} \geq 0$, ensuring the non-negativity of weights;
- $\sum_{j=1}^{n_i} w_{ij} = 1$, to normalize weights.

Optimal weights $\hat{w}_{ij}$, maximizing the profile EL under the given constraints, are found by the Lagrange multiplier method:

$$\hat{w}_{ij} = \frac{1}{n_i} \frac{1}{1 + \hat{\lambda}_i^\intercal (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i})}, \tag{3.4}$$

$$\text{where } \hat{\lambda}_i \text{ solves } \sum_{j=1}^{n_i} \frac{\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i}}{1 + \hat{\lambda}_i^\intercal (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i})} = 0.$$

3.2.3 Limitation of empirical likelihood and a best practice advice for SAE

The existence of an optimum solution to the maximization problem for the calibration weights $\hat{w}_{ij}$ is not necessarily guaranteed for applications in SAE. A necessary and sufficient condition ensuring the existence of a solution for $\hat{\lambda}_i$ is the existence of the zero vector as an interior point in the convex hull of constraint matrix $\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i}$. Especially for small sample sizes n_i this condition requires scrutiny (Emerson and Owen, 2009). If sample means of $\mathbf{x}_{ij}$ for area i strongly differ from $\bar{\mathbf{x}}_{\text{pop},i}$, for instance, due to a strong imbalance of individual sample values $\mathbf{x}_{ij}$ around the area-specific mean from population data $\bar{\mathbf{x}}_{\text{pop},i}$, no optimal solution for $\hat{\lambda}_i$ and subsequently $\hat{w}_{ij}$ can be obtained. The dimensionality of existing covariates p relative to the sample size n_i exacerbates the problem. As a result, the constraints in matrix $\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i}$ are infeasible for finding a global optimum in Equation (3.4). Concrete empirical examples are different largely unbalanced categorical covariates in $\mathbf{x}_{ij}$, leading to column-wise multicollinearity in the $n_i \times p$ matrix of constraints $\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i}$.

Overcoming mentioned technical requirements, Li et al. (2019) propose the use of the adjusted empirical likelihood (AEL) approach by Chen et al. (2008), which forces the existence of a solution to Equation (3.4). Essentially, the introduced adjustment is an additional pseudo-observation within each domain i, increasing area-specific sample sizes to n_{i+1}. This pseudo-observation is jointly calculated from respective area-specific survey and census means of covariates (Chen et al., 2008). Although the added adjustment-observation reduces risks of numerical instabilities, it simultaneously imposes difficulties from an applied perspective of SAE. Emerson and Owen (2009) scrutinize the application of AEL in the context of multivariate population means, maintaining that the added pseudo-observation distorts the true likelihood configuration even for moderate dimensions of p in cases of low area-specific sample sizes n_i. Chen et al. (2008, p. 430) note, that the problem is mitigated if the semi-parametric model is correctly specified and if the initial estimates for $\bar{\mathbf{x}}_{\text{smp},i}$ are not too far away from

the true population mean. Nevertheless, we observe that the influence of the bound-correction of Chen et al. (2008) used by Li et al. (2019) has drawbacks, which we will discuss in the model-based simulation in Section 3.4.

Dealing with empirical examples characterized by low domain-specific sample sizes, we abstain from the approaches of adding synthetic pseudo-observations to each domain. We maintain that in the context of non-linear semi-parametric approaches (such as RFs) there is a risk of including implausible individual predictions from f based on the pseudo-covariates, i.e. $\hat{y}_{\text{pseudo},i}$. In this sense, pseudo-observations manipulate the estimation of area-level means under limited auxiliary information in two ways: indirectly through their effect on the determination of all weights $\hat{w}_{ij}$ and directly through the predicted pseudo-value that is added to the survey sample.

We postulate a stepwise approach to ensure a solution to Equation (3.4) for each area i under a reduced risk of distortions driven by improper pseudo-values through optimization bound-corrections. This approach can be interpreted as a best-practice strategy on the incorporation of maximal auxiliary covariate information through calibration weights in Equation (3.4) for the estimation of area-level means with MERFs. In detail, we first check for each area i whether perfect column-wise-dependence in the $p \times n_i$ matrix of constraints $(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i})_{j=1,\ldots,n_i}$ exists. If so, we remove perfectly collinear columns and rerun the optimization. Subsequently, we proceed along two dimensions: a) increasing the sample size of i-th area and b) decreasing the number of auxiliary covariates p to calculate $\hat{w}_{ij}$ for area i. For a) we advise to sample a moderate number of observations (e.g., 10) randomly with replacement from an area which is "closest" to area i. We refer to areas as "closest", if they have the smallest Euclidean distance in census-level information $\bar{\mathbf{x}}_{\text{pop},i}$. This additionally allows to handle out-of-sample areas. For b) we propose a backward selection of covariate information based on the variable importance. Variable importance are RF-specific metrics that enable the ranking of covariates reflecting their influence on the predictive model. As we are primarily concerned about the order of influence of covariates, we rank based on the mean decrease in impurity importance, which measures the total decrease in node-specific variance of the response variable from splitting, averaged over all trees (Biau and Scornet, 2016). Overall, our strategy to handle potential failure in the solutions for weights and out-of-sample domains is summarized in the following algorithmic strategy:

---

1. Use MERF to obtain estimates $\hat{f}$, $\hat{u}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ from available unit-level survey data and estimate the indicator $\hat{\mu}_i^{\text{MERFagg}}$ (3.3) including weights $\hat{w}_{ij}$ following Equation (3.4).

2. If the calculation of weights fails due to infeasibility of constraints in the optimization problem for area i:

   (a) Check the feasibility of constraints used in the optimization and remove perfectly co-linear columns in $(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i})_{j=1,\ldots,n_i}$. Retry the optimization in Equation (3.4).

   (b) If the calculation of weights fails again, optionally enhance the domain-specific sample size of area i by sampling randomly with replacement from the most "sim-

ilar" domain according to the minimal row-wise Euclidean distance between area-specific aggregated covariate vectors $\bar{\mathbf{x}}_{\text{pop},i}$. Retry the calculation of weights $\hat{w}_{ij}$.

(c) If it fails again, reduce the number of covariates used for the calculation of weights for area i. Starting with the least influential covariate based on variable importance from $\hat{f}$, reduce the number of covariates in each step and retry the calculation of weights after each step.

(d) If the calculation of weights was not possible in step (c), set $\hat{w}_{ij}$ to $1/n_i$. These weights are non-informative for incorporating auxiliary information, however, the model-based estimates $\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i$ still comprise information from other in-sample areas.

3. Calculate the indicator for the i-th area as proposed by Equation (3.3).

The general performance is illustrated by the results of the model-based simulation in Section 3.4. Furthermore, the proposed best-practice strategy will be demonstrated in the application in Section 3.5.

3.3 Uncertainty estimation

The area-wise MSE is a conventional measure for SAE to assess the uncertainty of provided point estimates. While the quantification of uncertainty is essential for determining the quality of area-level estimates, its calculation remains a challenging task. For instance, even for the BHF model with block diagonal covariance matrices, the exact MSE cannot be analytically derived with estimated variance components (Prasad and Rao, 1990; Datta and Lahiri, 2000; González-Manteiga et al., 2008; Rao and Molina, 2015). Thus, the estimation of uncertainty by elaborate bootstrap-schemes is an established alternative (Hall and Maiti, 2006; González-Manteiga et al., 2008; Chambers and Chandra, 2013).

General statistical results concerning the inference of area-level indicators from MERFs in SAE are rare, especially in comparison to the existing theory of inference using LMMs. Although the theoretical background for predictions from RFs grows (Sexton and Laake, 2009; Wager et al., 2014; Wager and Athey, 2018; Athey et al., 2019; Zhang et al., 2019), existing research mainly aims to quantify the uncertainty of individual predictions. From a survey perspective, Dagdoug et al. (2022) recently analyse theoretical properties of RF in the context of complex survey data. The extension of these results for partly-analytical uncertainty measures in the context of dependent data structures and towards area-level indicators is non trivial and a conducive topic for theoretical SAE.

In this paper, we propose a non-parametric bootstrap for finite populations estimating the MSE of the introduced area-level estimator under limited aggregate information defined by Equation (3.3). Essentially, we aim to find a solution to two problems simultaneously: Firstly, we need to flexibly capture the dependence-structure of the data and uncertainty introduced by the estimation of Model (3.1). Secondly, we face problems in simulating a full bootstrap population in the presence of aggregated census-level data.

Our proposed solution to this dual problem is the effective combination of two existing bootstrap schemes introduced by Chambers and Chandra (2013) and González-Manteiga et al. (2008). Addressing the problem of non-parametric generation of random components, we rely on the approach introduced by Chambers and Chandra (2013). One key-advantage is its leniency to potential specification errors of the covariance structure, as the extraction of the empirical residuals only depends on the correct specification of the mean behaviour function f of the model. Solving the problem of missing unit-level population covariate data, we base the general procedure on the methodological principles of the parametric bootstrap for finite populations introduced by González-Manteiga et al. (2008) adapted to the estimation of domain-level means. This allows us to find (pseudo-)true values by generating only error components instead of simulating full bootstrap populations. An important step concerning the handling and resampling of empirical error components is centring and scaling them by a bias-adjusted residual variance proposed by Mendez and Lohr (2011). In short, the estimator of the residual variance under the MERF from Equation (3.2), $\hat{\sigma}^2_\epsilon$ is positively biased, as it includes excess uncertainty concerning the estimation of function $\hat{f}$. Further methodological details on the modification of the approach by Chambers and Chandra (2013) for MERFs for area-level means under unit-level models are found in Krennmair and Schmid (2022). Note that our proposed non-parametric MSE-bootstrap algorithm works for in- and out-of sample areas. The steps of the proposed bootstrap are as follows:

---

1. Use estimates $\hat{f}$, $\hat{\sigma}_e$, $\hat{\sigma}_u$, and respective weights $\hat{w}_{ij}$ from the application of the proposed method as summarized in Equation (3.3) on survey data with metric target variable y_{ij}.

2. Calculate marginal residuals $\hat{r}_{ij} = y_{ij} - \hat{f}(\mathbf{x}_{ij})$ and use them to compute level-2 residuals for each area by $\bar{r}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \hat{r}_{ij}$ for $i = 1, ...D$.

3. To replicate the hierarchical structure we use the marginal residuals and obtain the vector of level-1 residuals by $r_{ij} = \hat{r}_{ij} - \bar{r}_i$. Level-1 residuals r_{ij} are scaled to the bias-corrected variance $\hat{\sigma}^2_{\text{bc},\epsilon}$ (Mendez and Lohr, 2011) and centred, denoted by r^c_{ij}. Level-2 residuals $\bar{r}_i$ are also scaled to the estimated variance $\hat{\sigma}^2_v$ and centred, denoted by $\bar{r}^c$.

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

   (a) Simple random sampling with replacement (srswr) for each area i from the empirical distribution of scaled and centred level-1 (sample 1 value for each area i) and level-2 (sample n_i value for each area i) residuals to obtain the following three random components:

$$r^{*(b)}_{ij} = srswr(r^c_{ij}, n_i) \ , \ \bar{e}^{*(b)}_i = srswr(r^c_{ij}\frac{\hat{\sigma}_{\text{bc},\epsilon}}{\sqrt{N_i - n_i}}, 1), \text{ and}$$
$$u^{*(b)}_i = srswr(\bar{r}^c, 1).$$

   (b) Compute (pseudo-)true values for the population based on the fixed effects from

area-wise mean estimates $\hat{\mu}_i^{\text{MERFagg}}$, as:

$$\bar{y}_i^{(b)} = \sum_{j=1}^{n_i} \hat{w}_{ij} \hat{f}(\mathbf{x}_{ij}) + u_i^{*(b)} + \bar{E}_i^{(b)}, \qquad \text{where}$$
$$\bar{E}_i^{(b)} = \frac{n_i}{N_i} \bar{r}_{ij}^{*(b)} + \frac{N_i - n_i}{N_i} \bar{e}_i^{*(b)}.$$

(c) Use the known sample covariates $\mathbf{x}_{ij}$ to generate the bootstrap sample response values in the following way:

$$y_{ij}^{(b)} = \hat{f}^{\text{OOB}}(\mathbf{x}_{ij}) + u_i^{*(b)} + r_{ij}^{*(b)}.$$

We use OOB-predictions from $\hat{f}$ to imitate variations of $\mathbf{x}_{ij}$ covariates through predictions from unused observations within each tree in the fitting process that vary throughout the bootstrap replications.

(d) Estimate $\hat{\mu}_i^{\text{MERFagg}(b)}$ with the proposed method from Equation (3.3) on bootstrap sample values $y_{ij}^{(b)}$. Note that weights $\hat{w}_{ij}$ remain constant over B replications because the original survey covariates $\mathbf{x}_{ij}$ and population-level covariates $\bar{\mathbf{x}}_{\text{pop},i}$ remain unchanged over B.

5. Finally, calculate the estimated MSE for the area-level mean for areas $i = 1, ..., D$

$$\widehat{\text{MSE}}_i = \frac{1}{B} \sum_{b=1}^{B} \left[\left(\hat{\mu}_i^{\text{MERFagg}(b)} - \bar{y}_i^{(b)} \right)^2 \right].$$

3.4 Model-based simulation

The model-based simulation allows for a controlled empirical assessment of our proposed methods for point and uncertainty estimates. Overall, we aim to show, that the proposed methodology from Section 3.2 and Section 3.3 performs as well as traditional SAE methods and has advantages in terms of robustness against model-failure. In particular, we study the performance of the proposed MERFs under limited data access (*MERFagg*, (3.3)) to the *direct* estimator, the *TNER2* estimator proposed by Li et al. (2019), the *BHF* estimator (Battese et al., 1988) as well as the MERF assuming access to unit-level census data (*MERFind*, (3.2)) by Krennmair and Schmid (2022). The *direct* estimator only uses sampled data to estimate the mean, which implies a strong dependence between the area-specific sample size and the quality of estimates. The *BHF* model serves as an established baseline model for the estimation of area-level means under limited auxiliary data. The *TNER2* aims to provide an alternative to the *BHF*, introducing aspects of transformations under limited data access. General differences in the performance of the *direct*, *BHF*, and *TNER2* estimator to the two MERF candidates (*MERFagg*, *MERFind*) indicate advantages of semi-parametric and non-linear modelling in the given data scenarios. The additional inclusion of the *MERFind* enables a direct comparison regarding the effect of access to aggregated auxiliary data (*MERFagg*) and existing unit-level auxiliary data (*MERFind*).

Table 3.1: Model-based simulation scenarios

Scenario	Model	$x1$	$x2$	μ_i	v	ϵ
Normal	$y = 5000 - 500x_1 - 500x_2 + v + \epsilon$	$N(\mu_i, 3^2)$	$N(\mu_i, 3^2)$	$unif(-1, 1)$	$N(0, 500^2)$	$N(0, 1000^2)$
Pareto	$y = 5000 - 500x_1 - 500x_2 + v + \epsilon$	$N(\mu_i, 3^2)$	$N(\mu_i, 3^2)$	$unif(-1, 1)$	$N(0, 500^2)$	$Par(3, 800)$
Interaction	$y = 1000 + 100x_1x_2 + 75x_2 + v + \epsilon$	$N(\mu_i, 2^2)$	$N(\mu_i, 1)$	$unif(-7, 7)$	$N(0, 500^2)$	$N(0, 1000^2)$
Logscale	$y = \exp(7.5 - 0.25x_1 - 0.25x_2 + v + \epsilon)$	$N(\mu_i, 1)$	$N(\mu_i, 1)$	$unif(-3, 3)$	$N(0, 0.15^2)$	$N(0, 0.25^2)$

We consider four scenarios denoted as *Normal*, *Pareto*, *Interaction*, and *Logscale* and repeat each scenario independently $M = 500$ times. All four scenarios assume a finite population U of size $N = 50000$ with $D = 50$ disjunct areas $U_1, ..., U_D$ of equal size $N_i = 1000$. We generate samples under stratified random sampling, utilizing the 50 small areas as stratas, resulting in a sample size of $n = \sum_{i=1}^{D} n_i = 1229$. The area-specific sample sizes range from 5 to 50 sampled units with a median of 21 and a mean of 25. The sample sizes are comparable to area-level sample sizes in the application in Section 3.5 and can thus be considered to be realistic.

The choice of the simulation scenarios is motivated by our aim to evaluate the performance of the competing methods for economic and social inequality data. This includes skewed data, deviations from normality of error terms, or the presence of unknown non-linear interactions between covariates, that might trigger model-misspecifications in traditional SAE approaches based on LMMs. The data generating processes for the used scenarios are provided in Table 3.1. Scenario *Normal* provides a baseline under a LMM with normally distributed random effects and unit-level errors. As the model assumptions for LMMs are fully met, we aim to show that the *MERFagg* performs similarly well compared to linear competitors. Scenario *Pareto* is based on the same linear additive structure as scenario *Normal*, but has Pareto distributed unit-level errors. This leads to a skewed target variable, comparable to empirical cases of monetary data. The data generating process of scenario *Interaction* likewise results in a skewed target variable y_{ij}, although it shares its structure of random components with *Normal*. The *Interaction* scenario portrays advantages of semi-parametric and non-linear modelling methods protecting against model-failure arising from models with unknown interactions. Scenario *Logscale* introduces an additional example resulting in a skewed target variable. Log-normal distributed variables mimic realistic income scenarios and constitute a showcase for SAE transformation approaches. We want to show the ability of MERFs and particularly of *MERFagg* to handle such scenarios as well by identifying the non-linear relation introduced trough the transformation on the linear additive terms.

We evaluate point estimates for the area-level mean over M replications by the empirical root MSE (RMSE), the relative bias (RB), and the relative root mean squared error (RRMSE). As quality-criteria for the evaluation of the MSE estimates, we choose the relative bias of RMSE (RB-RMSE) and the relative root mean squared error of the RMSE (RRMSE-RMSE):

$$\text{RMSE}_i = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\mu}_i^{(m)} - \mu_i^{(m)})^2},$$

$$\text{RB}_i = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{\mu}_i^{(m)} - \mu_i^{(m)}}{\mu_i^{(m)}}\right),$$

$$\text{RRMSE}_i = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{\mu}_i^{(m)} - \mu_i^{(m)}}{\mu_i^{(m)}}\right)^2},$$

$$\text{RB-RMSE}_i = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M} MSE_{\text{est},i}^{(m)}} - RMSE_i}{RMSE_i},$$

$$\text{RRMSE-RMSE}_i = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\sqrt{MSE_{\text{est},i}^{(m)}} - RMSE_i\right)^2}}{RMSE_i},$$

where $\hat{\mu}_i^{(m)}$ is the estimated mean in area i based on any of the methods mentioned above and $\mu_i^{(m)}$ defines the true mean for area i in replication m. $MSE_{\text{est},i}^{(m)}$ is estimated by the proposed bootstrap from Section 3.3.

For the computational realization of the model-based simulation, we use R (R Core Team, 2022). The *BHF* estimates are realized from the **sae**-package (Molina and Marhuenda, 2015). For the estimates of the *TNER2*, we used code provided by Li et al. (2019). For estimates based on the MERF approach, we use the packages **ranger** (Wright and Ziegler, 2017) and **lme4** (Bates et al., 2015) to implement our method (*MERFagg*) and the *MERFind* estimator (Krennmair and Schmid, 2022). For RFs, we set the number of split-candidates to 1, keeping the default of 500 trees for each forest.

3.4.1 Performance of point estimators of the small area means

We start with a focus on the performance of point estimates. Figure 3.1 reports the empirical RMSE of each point estimation method under the four scenarios. As expected, the *direct* estimates perform poorest due to the low sample sizes and the complexity of the data generating process. In these specific settings, the *TNER2* estimator outperforms *direct* estimates but performs worse compared to the *BHF*. In the *Pareto* and *Logscale* scenario, benefits of transformations might be suppressed by the influence of pseudo-observations due to the AEL approach, as discussed throughout the methodological Section 3.2.3 of this paper.

In the *Normal* scenario, the *BHF* performs best as it replicates the data generating process. The *MERFind* and the *MERFagg* perform on a comparable level, underlining the quality of our proposed calibration approach to incorporate aggregated census-level information through the weights. *MERFagg* shows a better performance in median values, however the range of area-specific RMSE values is larger compared to MERF estimates based on unit-level census information. One area with particularly low sample size has a relatively high level of RMSE, which is explainable by the dependence of the optimum function for the weights in Equation (3.4) on n_i.

We observe similar patterns in the *Pareto* scenario. The *BHF* has one outlier for an area with low sample size. As anticipated, the performance of both MERF candidates is comparable to the *Normal* scenario, confirming robust behaviour under skewed data and violations of the normal distribution of errors. Since *MERFagg* behaves comparably, the robustness also holds for the calculation of calibration weights.

In the *Interaction* scenario, the point estimates of the proposed *MERFagg* outperform tra-

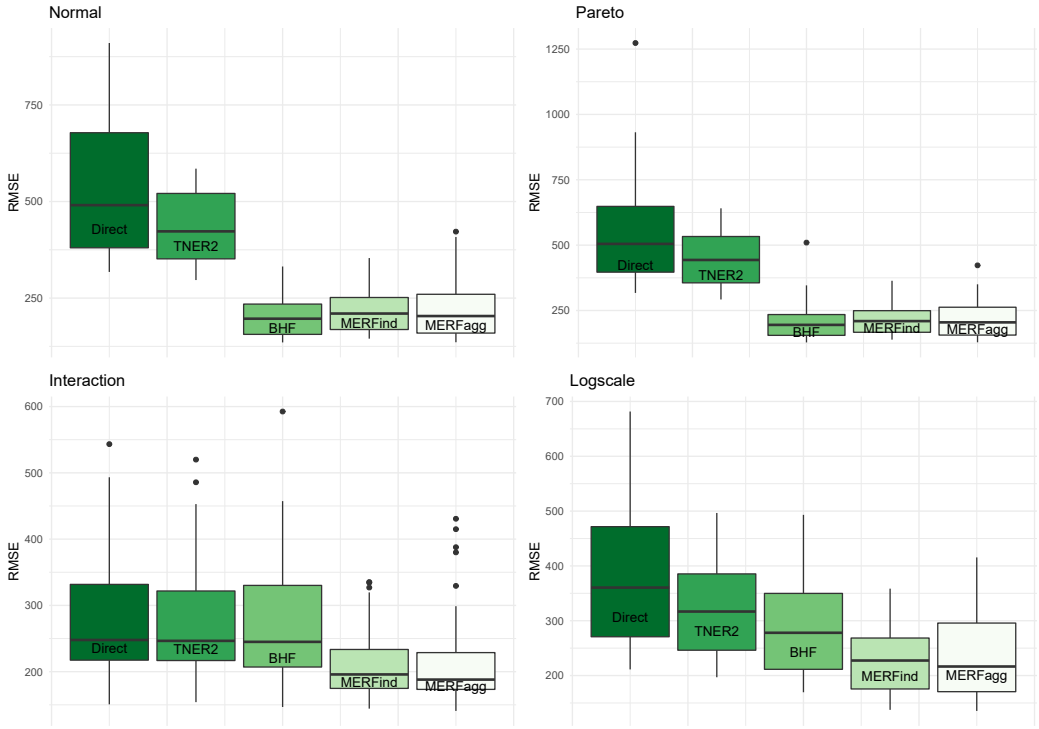


Figure 3.1: Empirical RMSE comparison of point estimates for area-level averages under four scenarios

ditional SAE approaches under limited auxiliary information. Apparently the LMM-based methods cannot sufficiently capture the underlying predictive relation between the covariates, while the MERFs detect the non-linear term. Regarding the impact of restricted covariate data access, we observe relatively low values of mean and median RMSE compared to the hypothetical case of existing unit-level data in *MERFind*. Four outliers in areas with low sample sizes for *MERFagg* become apparent, although the median RMSE is lowest. We maintain, that this phenomenon can be mitigated if we increase the size of "close" observations from other areas to a higher level, especially in cases of complex interactions of effects in covariates such as *Interaction*.

The last scenario *Logscale* shows that the *MERFagg* outperforms the *direct* and LMM-based competitors. Similar to the *Interaction* and *Pareto* scenario, the effect of covariate data access - comparing *MERFagg* and *MERFind* - is not severe for an average area.

Overall, the results from Figure 3.1 indicate that the MERF performs comparably well to LMMs in simple scenarios, and outperforms traditional SAE models in the presence of complex data generating processes, such as unknown non-linear relations between covariates or non-linear functions. Additionally, the robustness against model-misspecification of MERFs and their calibration weights $\hat{w}_{ij}$ holds if distributional assumptions for LMMs are not met, i.e. in the presence of non-normally distributed errors and skewed data. The influence of unit-level versus aggregated covariate information appears to be marginal in all of our four scenarios. We observe a moderate dependence between sample sizes and the quality of area-specific means for *MERFagg*, which is mainly explained by the way the calibration weights rely on the quality of survey data for a respective area i as discussed in Section 3.2.2.

Table 3.2: Mean and Median of RB and RRMSE over areas for point estimates in four scenarios

	Normal		*Pareto*		*Interaction*		*Logscale*	
	Median	Mean	Median	Mean	Median	Mean	Median	Mean
RB								
Direct	0.0000	0.0002	0.0001	0.0004	-0.0005	0.0076	0.0003	0.0010
TNER2	0.0002	-0.0001	-0.0003	-0.0008	0.0010	0.0187	-0.0014	-0.0020
BHF	0.0009	0.0013	0.0019	0.0022	0.0031	0.0233	-0.0188	-0.0225
MERFind	0.0014	0.0019	0.0033	0.0038	0.0071	0.0061	0.0076	0.0082
MERFagg	0.0001	0.0005	0.0011	0.0016	0.0034	0.0138	0.0004	0.0002
RRMSE								
Direct	0.0984	0.1080	0.0994	0.1100	0.1570	1.1500	0.0978	0.1030
TNER2	0.0838	0.0886	0.0876	0.0915	0.1550	1.2900	0.0866	0.0879
BHF	0.0392	0.0418	0.0368	0.0418	0.1590	1.2900	0.1670	0.1760
MERFind	0.0417	0.0450	0.0398	0.0441	0.1370	1.5900	0.0620	0.0636
MERFagg	0.0409	0.0451	0.0409	0.0446	0.1330	1.2900	0.0610	0.0634

Table 3.2 reports the corresponding values of RB and RRMSE for the discussed point estimates. The RB and the RRMSE from the *MERFagg* attest a competitively low level under all scenarios. All model-based MERF estimators have a lower mean and median RRMSE compared to the *direct* estimator in all scenarios. Despite a few outliers for RMSE and RB (cf. Figure 3.1), the median and mean values of *MERFagg* are remarkably low emphasizing the quality of estimates given the the substantial reduction in required covariate information.

3.4.2 Performance of the bootstrap MSE estimator

We scrutinize the performance of our proposed MSE estimator on the four scenarios, examining whether the proposed procedure for uncertainty estimates performs equally well in terms of robustness against model-misspecification and in cases of limited access to auxiliary information.

For each scenario and each simulation round, we choose $B = 200$ bootstrap replications. From the comparison of RB-RMSE among the four scenarios provided in Table 3.3, we infer, that the proposed non-parametric bootstrap-procedure effectively handles all four scenarios. This is demonstrated by relatively low mean values of positive RB-RMSE over the 50 areas after M replications. From an applied perspective, we prefer over- to underestimation for the MSE as it serves as an upper bound. We mainly use the area-level MSE for the further assessment in terms of CVs and consequently overestimation of area-level MSEs leads to an increased CVs. If our CVs are still below the thresholds, the estimates are definitely acceptable. The difference in RB-RMSE between *Normal* and *Pareto* is marginal, indicating that the non-parametric bootstrap effectively handles non-Gaussian error terms.

Figure 3.2 provides additional intuition on the quality of our proposed non-parametric MSE-bootstrap estimator. Given the area-wise tracking properties in all four scenarios, we conclude that our MSE estimates strongly correspond to the empirical RMSE. We infer that the overestimation in Table 3.3 is mainly driven by overestimation in areas with low sample sizes. Thus, our non-parametric MSE estimator provides an upper bound for the uncertainty of

Table 3.3: Performance of MSE estimator in model-based simulation: mean and median of RB-RMSE and RRMSE-RMSE over areas

	Normal		*Pareto*		*Interaction*		*Logscale*	
	Median	Mean	Median	Mean	Median	Mean	Median	Mean
RB-RMSE	0.0525	0.0591	0.0596	0.0643	0.0192	0.0205	-0.0117	0.0054
RRMSE-RMSE	12.7000	15.6000	30.6000	34.3000	9.9000	12.4000	22.9000	25.3000

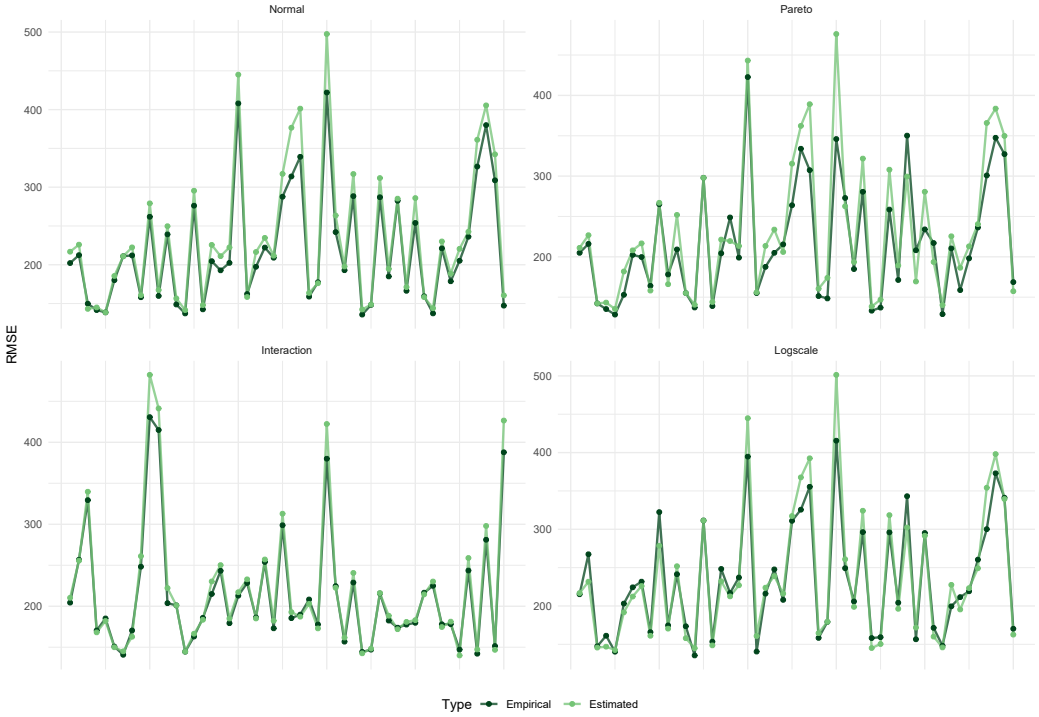


Figure 3.2: Estimated and empirical area-level RMSEs for four scenarios

particular difficult point estimates due to low sample sizes. Apart from this characteristic, we observe no further systematic differences between the estimated and empirical MSE estimates regarding their performance throughout our model-based simulation.

3.5 Application

This section starts with a description of data sources and outlines our empirical analysis. We describe the survey data SOEP (Socio-Economic Panel) and discuss primary *direct* estimates on spatial differences of average individual opportunity cost of care work for German RPRs. Moreover, we propose the use of model-based SAE, which incorporates auxiliary variables from the 2011 German census. Demonstrating our proposed method of MERFs with aggregated data for point and uncertainty estimates, we show advantages to existing model-based SAE methods. Finally, we discuss our empirical findings concerning the cost of care work in Germany. We conduct the analysis with R (R Core Team, 2022).

3.5.1 Data sources and direct estimates of spatial opportunity cost of care work

The SOEP was established in 1984 by the German Institute of Economic Research (DIW) and evolved into an imperative survey for Germany regarding multidisciplinary social information on private households (Goebel et al., 2019). Statistical considerations regarding sampling designs and representativeness of the longitudinal data set, justify its relevance for governmental institutions, policy makers, and researchers alike. For our primary calculation of opportunity cost of care work, we need information on individual income as well as hours worked on the job and for care work. This information is only provided in the SOEP, in contrast to the German Microcensus (Statistisches Bundesamt, 2015), where income is only available as an interval censored variable.

We construct the target variable of individual monthly opportunity cost of care work from the SOEP in 2011 (Socio-Economic Panel, 2019) and use the available refreshment samples. We choose the year 2011 because the last census was in this year and therefore census and survey data have no time inconsistencies. The underlying sampling design is a multi-stage stratified sampling procedure: Initially, stratification is carried out into federal states, governmental regions, and municipalities. Subsequently, addresses are sampled using the random walk methodology within each primary sampling unit (Kroh et al., 2018). Our analysis focuses on the working age population aged between 15 to 64, as defined by international standards (OECD, 2020). In detail, we calculate the individual opportunity cost in Euro per month for 2011 as follows: first, we compute opportunity cost as hourly wage by taking the mean gross individual income divided by hours of paid work. Then, we multiply the hours of monthly unpaid work due to child- or elderly-care by the hourly cost of opportunity. The resulting metric target variable y_{ij} for Germany is highly skewed, ranging from 0€ to 2413.79€ (mean: 100.96€ and median: 176.93€). A histogram is provided in Figure 3.3.

In total we have 3939 sample survey observations. National averages do not serve for monitoring efficacy of regional developments and policy measures. Our major interest is a finer spatial resolution to map regional patterns of opportunity cost of care work across Germany. We analyse 96 respective RPRs in Germany, resulting in area-specific sample sizes from 4 to 158 with a mean of 35 and median of 41. First results of *direct* estimates can be seen in the map in Figure (3.3). Estimates of the mean monthly opportunity cost of individual care work range from 64.31€ (Oberpfalz-Nord) to 409.38€ (Neckar-Alb). In general, we observe no major difference between former East and West Germany. Additionally, levels of opportunity cost are higher in metropolitan areas surrounding cities than in the cities itself and compared to rural areas.

Small sample sizes lead to unreliable estimates accompanied by high variances. Furthermore, we are not allowed to report *direct* estimates from regions with sample size below 10 due to confidentiality agreements with the data provider. This is the case for 7 RPRs. To obtain variances and subsequently determining the coefficients of variation (CV) for the *direct* estimates, we use the calibrated bootstrap by Alfons and Templ (2013) implemented in the R-package **emdi** by Kreutzmann et al. (2019). Eurostat (2019a) postulates that estimates with a CV of less than 20% can be considered as reliable. As reported by Figure 3.3, more than half of the regions (47 out-of reaming 89) exceed this threshold.

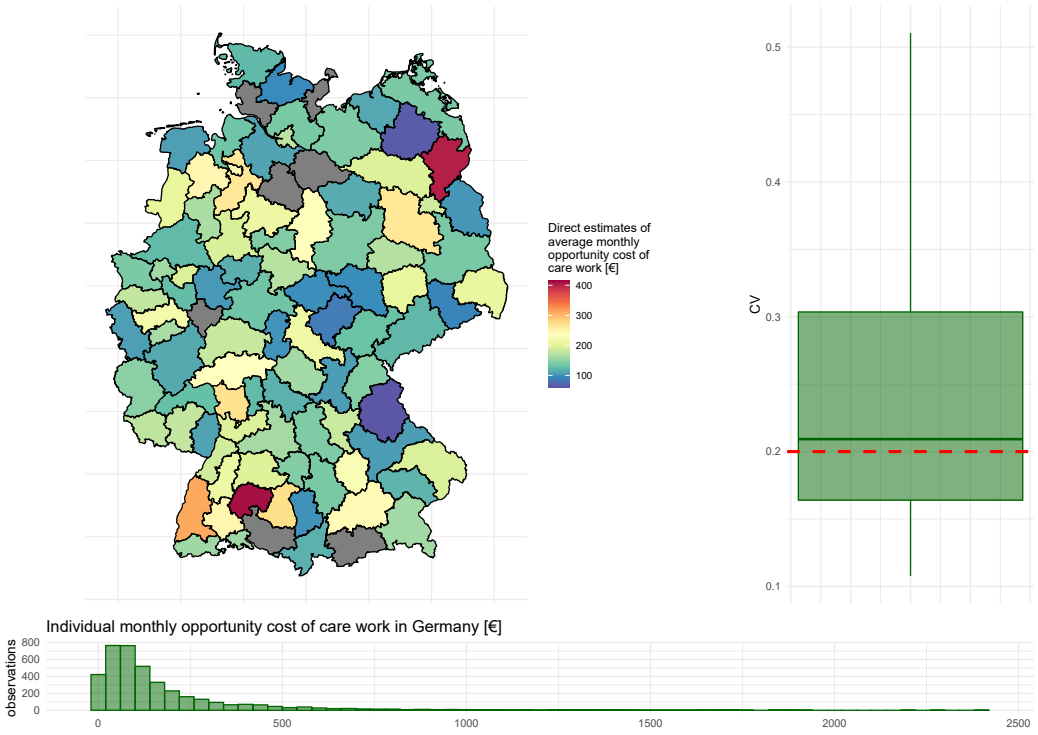


Figure 3.3: Overview of *direct* estimates, corresponding CVs and the distribution of opportunity cost of care work in Germany.

The *direct* estimation results suffer from differences in quality due to low area-level sample sizes and specifically high variability. Model-based SAE methods help to improve the estimation accuracy of results. As SOEP auxiliary variables are measured in the same way as in the Germans census (Statistisches Bundesamt, 2015), census covariate data can serve as auxiliary information needed in SAE models. However, the German census provides information only at aggregated RPR-levels. Overall, we have 19 covariates on personal and socio-economic background within our sample for which we additionally received corresponding means from the German Statistical Office calculated from the German 2011 census. Details on available covariates and their variable importance is provided within the Appendix in Table B.1.

3.5.2 Model-based estimates

This section illustrates the application of our proposed method for MERFs with aggregate covariate data for the estimation of area-level means. We map the estimated monthly mean opportunity cost of unpaid care work for 96 RPRs in Germany for the year 2011. Moreover, we assess the quality of our estimates by providing CVs based on our proposed non-parametric MSE-bootstrap procedure discussed in Section 3.3 and juxtapose our results to the previously discussed *direct* estimates and the well-established BHF model by Battese et al. (1988). A full comparison to the *TNER2* estimates (Li et al., 2019) is not possible because Li et al. (2019) do not provide uncertainty estimators required for a qualitative comparison in terms of CVs.

As reported by Figure 3.3, our target variable of individual opportunity cost is highly skewed, indicating that traditional LMMs (such as the BHF) run the risk of model-misspecification. In contrast, our proposed procedure shows robustness against model-failure due to

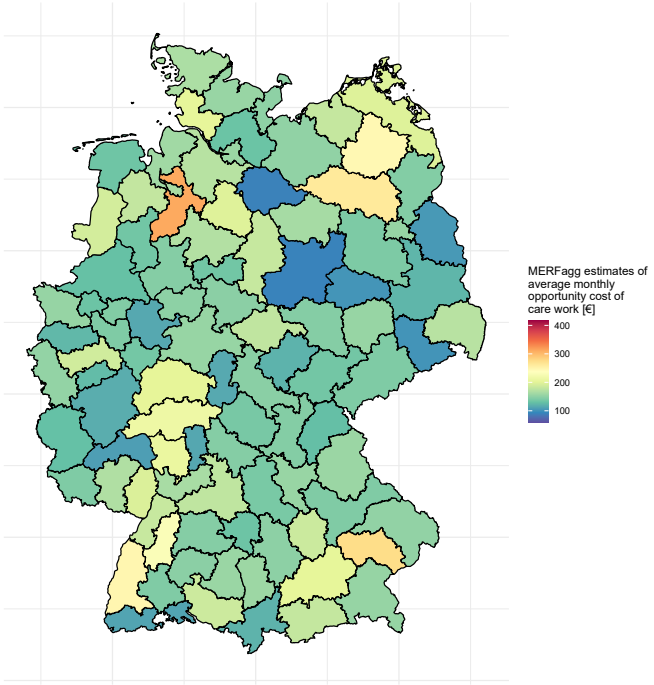


Figure 3.4: Spatial representation of area-level mean estimates from *MERFagg* (3.3) for mean monthly opportunity cost of care work [€].

outliers or complex data structures. Apart from specifying separate regions being modelled as random intercepts, the proposed *MERFagg* approach can be seen as purely data-driven: We train a predictive model on the survey set and incorporate as much auxiliary information for the determination of area-specific calibrations weights as possible based on the variable importance obtained from the fitted RF object $\hat{f}$. For this example we set the tuning parameter of the RF to 500 sub-trees. Repeated 5-fold cross-validation supports the choice of proposing 5 randomly drawn split candidates at each split for the forest. Regarding our best-practice strategy, we chose that we want to calculate the weights based on a minimum of the 3 most influential variables. An overview of the number of covariates included can be found in the appendix (Figure B.1). For the non-parametric MSE bootstrap-procedure, we use $B = 200$.

The results from the application of *MERFagg* are reported in Figure 3.6. We primarily focus on a discussion of technical details of estimates from our proposed approach and postpone the contextual discussion of results to the end of this section. Overall we observe a dominance of covariates of age, size of the household, households with a child, gender and whether the person is employed in the public sector (cf. Table B.1 in the Appendix). Throughout all 96 areas, we incorporate auxiliary information from 3 up to 15 covariates from census-level aggregates through optimal calibration-weights $\hat{w}_{ij}$. A detailed map on the number of included census-level covariates is provided in the Appendix within Figure B.1. Unfortunately this attempt failed for 5 regions, which were left with uninformative weights $\hat{w}_{ij} = 1/n_i$. Although these estimates do not incorporate auxiliary information, recall from Equation (3.3) that the corresponding estimates are reduced to $\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i$ and thus still rely on the model-based estimates comprising information from other in-sample areas.

A comparison between the maps from *direct* estimates in Figure 3.3 and estimates based

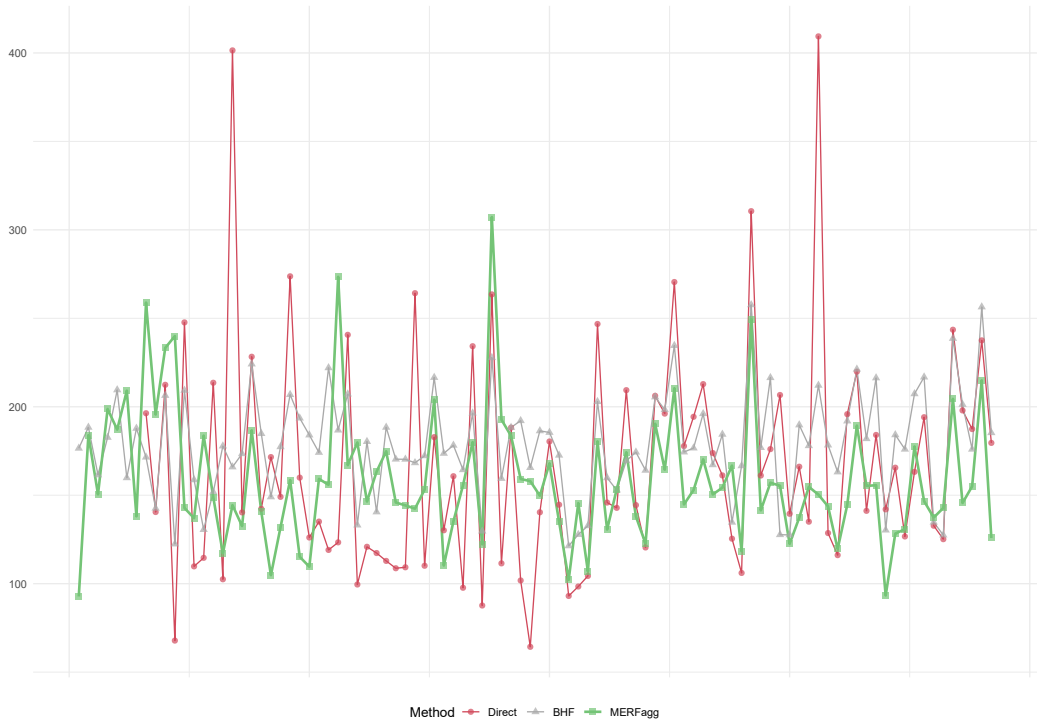


Figure 3.5: Detailed comparison of area-level mean estimates for monthly opportunity cost of care work [€]. The 96 German RPRs are sorted by increasing sample size. We compare results based on methods *direct*, *BHF*, and *MERFagg*.

on *MERFagg* from Figure 3.4 indicates that results form *MERFagg* appear to be more balanced and overall no major differences regarding changes in regional patterns of opportunity cost of care work are observable. Figure 3.5 sorts areas by increasing survey sample sizes and thus allows for a more precise discussion on peculiarities of point estimates for area-level means of monthly opportunity cost for the 96 RPRs. Estimates from the BHF method are produced form the R-package **sae** (Molina and Marhuenda, 2015). Although, the raw comparison of point estimates only allows for limited findings regarding the quality of methods, we report the mitigation of two outlier-driven direct estimates. Compared to the *direct* estimates, as well as the estimates from the BHF, the *MERFagg* produces relatively lower values although the estimates track patterns of high- and low levels with increasing survey sampling size.

As already discussed, *direct* estimates suffer from relatively low accuracy measured by their respective CVs. Figure 3.6 juxtaposes CVs for *direct* estimates, the *BHF*, and our proposed method of *MERFagg* to contextualize the performance of point estimates from Figure 3.5. We observe that CVs for *MERFagg* are on average smaller compared to CVs from *direct* estimates as well as the BHF. According to the boxplots in Figure 3.6, model-based estimates produce more accurate results indicated by lower CVs than *direct* estimates. *MERFagg* shows the lowest CVs compared to the other methods in mean and median-terms. Two areas can be considered as outliers reporting CVs over 0.3. For one of these two regions, the calculation of weights failed. The *MERFagg* estimates improve the *direct* estimates: Only 15 areas from 96 do not meet the required threshold of 20%. As expected, especially for areas that are unreliable due to low sample sizes, model-based estimates improve the accuracy. In turn, we observe

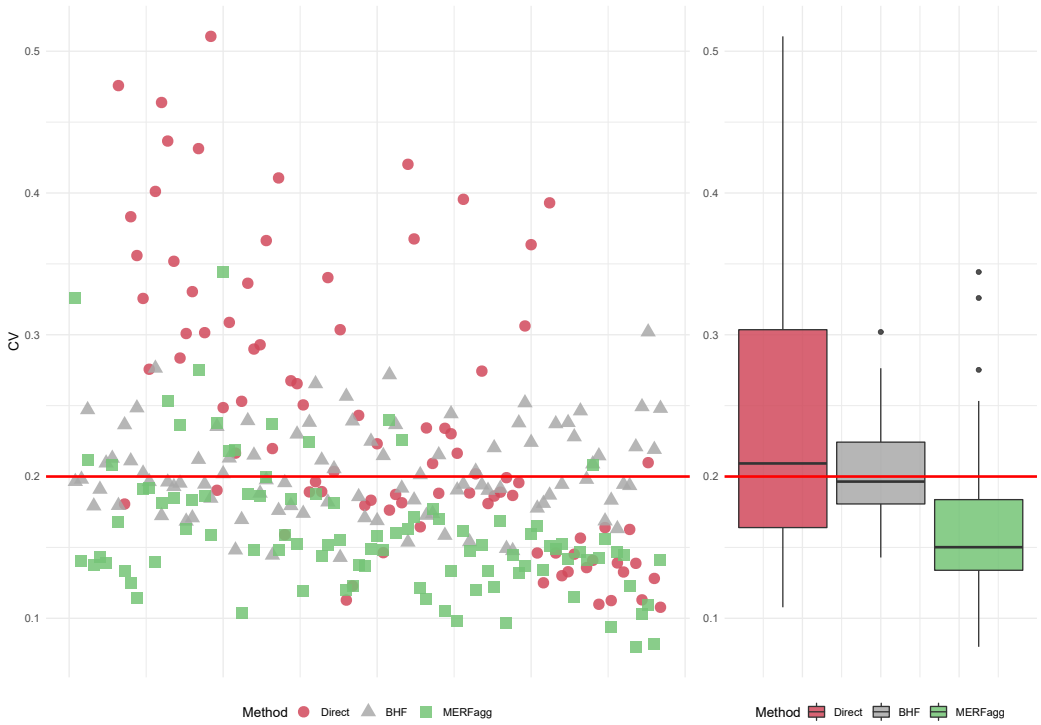


Figure 3.6: Left: Comparison of area-specific CVs ordered from low to high sample sizes. Right: Comparison of CVs over 96 respective areas between direct, *BHF* and *MERFagg*. The red line marks the 20%-criterion for defining reliable estimates by Eurostat (2019a).

that the *direct* estimates are relatively accurate for areas with high sample sizes. Compared to other model-based SAE methods, survey weights are not directly used in the model-fitting for *MERFagg*. Although it is generally possible to incorporate survey weights in the importance sampling within a forest, we maintain that the efficient use of survey weights with MERFs for the estimation of area-level indicators requires further research which would exceed the scope of this paper.

Overall, all RPRs throughout Germany report comparable levels of average individual monthly opportunity cost of care work. Nevertheless, a detailed inspection of Figure 3.4 reveals a small cluster of lower values in the North-East of Germany. From a causal perspective, the explanation of such patterns appears to be difficult and not effective. Wage and individual opportunity cost directly relate while time spent for care work negatively affects opportunity cost. Thus, it is not observable whether the effect is driven by differences in average income or increased time-allocation for care work or both. On the other hand, the concept allows us to uncover and map the value of unpaid care work on a sub-regional-level in Germany.

3.6 Conclusion

In this paper, we provide a coherent framework enabling the use of RFs for SAE under limited auxiliary data. Our approach meets modern requirements of SAE, including the robustness against model-failure and aspects of data-driven model-selection within the existing methodological framework of SAE. We introduce a semi-parametric unit-level mixed model, treating

LMM-based SAE methods, such as the BHF and the EBP, as special cases. Furthermore, we discuss the MERF procedure (Hajjem et al., 2014) and its application to SAE as introduced by Krennmair and Schmid (2022). We address the challenging task of incorporating aggregated census-level auxiliary information for MERFs and propose the use of calibration weights based on a profile EL optimization problem. We deal with potential issues of numerical instabilities of the EL approach and propose a best practice strategy for the application of our proposed estimator *MERFagg* for SAE. The proposed point estimator for area-level means is complemented by a non-parametric MSE-bootstrap-scheme. We evaluate the performance of point and MSE estimates compared to traditional SAE methods by a model-based simulation that reflects properties of real data (e.g., skewness). From these results, we conclude that our approach outperforms traditional methods in the existence of non-linear interactions between covariates and demonstrates robustness against distributional violations of normality for the random effects and for the unit-level error terms. Moreover, we observe that the inclusion of aggregated information through calibration weights based on EL works reliably. Regarding the performance of our MSE-bootstrap scheme, we observe moderate levels of overestimation and report authentic tracking behaviour between estimated and empirical MSEs. We focus on a distinctive SAE example, where we study the average individual opportunity cost of care work for Germany RPRs. Overall, we provide an illustrative example on how to use our data-driven best practice strategy on MERFs in the context of limited auxiliary data. Comparing direct to model-based results, we show that differences between German RPRs are small and balanced. Nevertheless, we allocate a small cluster of lower levels of average individual opportunity cost of care work in the North-Eastern part of Germany.

From an empirical perspective, we face limitations that directly motivate further research. Firstly, we only calculate the opportunity cost of the working population and neglect care work done by people who already left the labour market due to care work issues. Despite its long tradition in economics, the basic concept of opportunity cost (treating the shadow value of care work equivalently to hourly wage from labour) faces drawbacks. Different models from a health and labour economic perspective (e.g., Oliva-Moreno et al. (2019)) can be integrated into our approach. Nevertheless, given the data and our initial aim to provide a general methodology for regional mapping of care work specific regional differences, we consider the hourly wage as a first reasonable approximation to the unobservable "real" shadow price.

We motivate two major dimensions for further research, including theoretical work and aspects of generalizations. From a theoretical perspective, further research is needed to investigate the construction of a partial-analytical MSE for area-level means or the construction of an asymptotic MSE estimator. From a statistical perspective, an in-depth analysis regarding the effects of incorporating survey weights into RFs and particularly MERFs under aggregated covariate data is needed for point and uncertainty estimates, as this would clearly exceed the scope of the present paper. Our approach shares the EL-calibration-argument with Li et al. (2019), however, saves on the computationally intensive procedure of a smearing step (Duan, 1983) without drawbacks on the predictive performance, because no transformations and corresponding bias exists. Nevertheless, we maintain that pairing our approach with a smearing argument allows for a more general methodology and subsequently for the estimation of in-

dicators such as quantiles (Chambers and Dunstan, 1986). Although, we will leave a detailed discussion of this idea to further research, a short outline of the argument can be found in the Appendix B.2. Apart from generalizations to quantiles, the approach of this paper is generalizable to model (complex) spatial correlations. Additionally, a generalization towards binary or count data is possible and left to further research. The semi-parametric composite formulation of Model (3.1) allows for f to adapt any functional form regarding the estimation of the conditional mean of y_{ij} given x_{ij} and technically transfers to other machine learning methods, such as gradient-boosted trees or support vector machines.

Acknowledgements

Würz gratefully acknowledges support by a scholarship of Studienstiftung des deutschen Volkes. The authors are grateful for the computation time provided by the HPC service of the Freie Universität Berlin.

Appendix B

B.1 Additional information on the application (Section 3.5)

Table B.1: Auxiliary variables on personal and socio-economic background and their variable importance based on the trained RF $\hat{f}$.

Covariates	**Variable importance**
Age in years	30715147.623
Number of persons living in household	17109846.300
Position in Household: Child	7519805.884
Sex	4031803.086
Employment status: civil servants	3704520.439
Employment status: employed without national insurance (e.g. mini-jobber)	3078656.890
Tenant or owner	2632970.858
Position in Household: single parent	2500261.812
Migration background: direct	2453187.125
Position in Household: living alone	1380917.681
Position in Household: marriage-like	1341933.482
Migration background: indirect	1207604.491
Grouped nationality: European Union (excluding Germany)	697919.972
Grouped nationality: remaining European countries	468653.092
Grouped nationality: Asia	367207.174
Grouped nationality: North America	224042.331
Grouped nationality: Australia	45084.788
Grouped nationality: Africa	10109.844
Grouped nationality: South America	5150.957

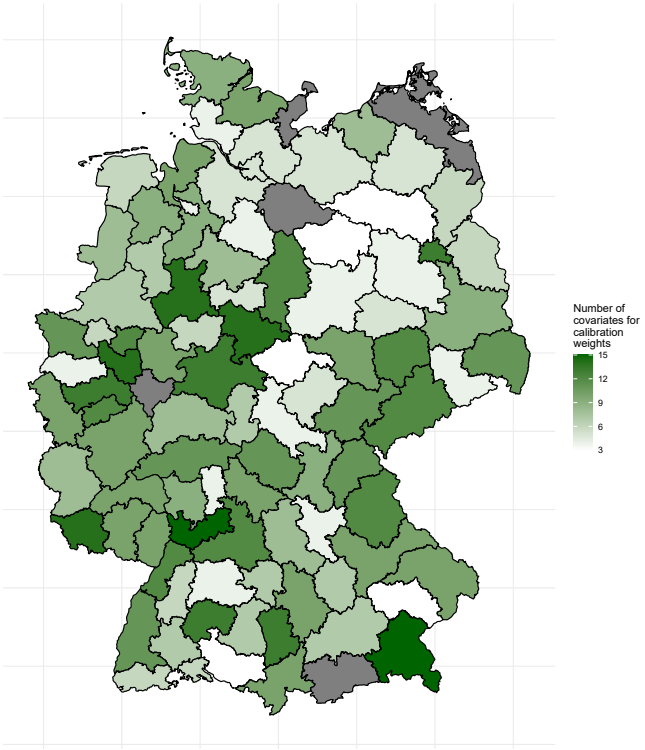


Figure B.1: Inclusion of covariates through weights

B.2 Extension towards the estimation of quantiles

Smearing approach and estimation of means: The smearing argument form Duan (1983) could be optionally inserted in Equation (3.3) to estimate mean values

$$\hat{\mu}_i^{\text{MERFagg Smearing}} = \sum_{j=1}^{n_i} \left[\hat{w}_{ij} \frac{1}{R} \sum_{r=1}^{R} (f(\mathbf{x}_{ij}) + \hat{u}_i + e_{ir}^*) \right], \tag{B.1}$$

where R is a suitably large number of smearing residuals and e_{ir}^* are OOB model residuals:

$$e_{ij}^* = y_{ij} - f(\mathbf{x}_{ij})^{\text{OOB}} - \hat{u}_i.$$

Note that the formulation of Equation (B.1) coincidences with the estimator of Li et al. (2019), if we choose $f = \mathbf{x}_{ij}^\intercal \beta$ and draw e_r^* from $N(0, \hat{\sigma}_e^2)$. Additionally, they apply a data-driven transformation on $f(\mathbf{x}_{ij}) + \hat{u}_i + e_{ir}^*$.

Extension towards quantile estimation: The combination of a smearing argument (Duan, 1983) with a model of a finite-population CDF of y enables the estimation of area-specific CDFs for y_i. Chambers and Dunstan (1986) develop a model-consistent estimator for a finite-population CDF from survey data and provide asymptotic results under LMMs. Tzavidis et al. (2010) propose the use of the CDF method within a general unit-level SAE framework to produce estimates of means and quantiles using robust methods. In the case of RF, it holds that the predicted value of a non-sampled individual observation in area i is given by $\hat{\mu}_{ij} = \hat{f}(\mathbf{x}_{ij}) + \hat{u}_i$, which expresses its expected value conditional on area i. We propose to obtain an estimator of the area-level CDF $\hat{F}_i^*(t)$ using existing survey information modifying the CDF method, by substituting $\hat{\mu}_{ij} = \hat{f}(\mathbf{x}_{ij}) + \hat{u}_i$ and incorporating census-level information for unsampled predictions via weights $\hat{w}_{ij}$. The respective estimator for the area-level CDF $\hat{F}_i^*(t)$ is summarized as:

$$\hat{F}_i^*(t) = N_i^{-1} \left[\sum_{j \in s_i} I(y_{ij} \leq t) + R^{-1} \sum_{j \in s_i} \sum_{r=1}^{R} n_i \hat{w}_{ij} I\left(\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i + e_{ir}^* \leq t \right) \right], \tag{B.2}$$

where $e_{ij}^* = y_{ij} - f(\mathbf{x}_{ij})^{\text{OOB}} - \hat{u}_i$. The area-level quantile $q(i, \phi)$ of $\phi \in [0, 1]$ can straight forwardly be calculated by:

$$\hat{q}_i(\phi) = \hat{F}_i^{*-1}(\phi).$$

Part II

Estimation of Regional Economic Indicators from Area-Level Models

Chapter 4

Experimental UK regional consumer price inflation with model-based expenditure weights

Chapter 5

Estimating regional unemployment with mobile network data for functional urban areas in Germany

5.1 Introduction

Since jobs are predominantly located in cities, more people move to the cities. For example, the continuous growth of cities is creating shortages on the German housing and real estate markets (Möbert, 2018). Most large cities have higher population growth rates than the national average (see e.g., an interactive map of the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR, 2017)). Due to urban labour migration, the number of people living in cities is steadily increasing nationwide. As Buch et al. (2014), smaller cities recorded less net immigration than large cities, which is caused by the attractiveness of larger cities and the advantages of living in them. These are better infrastructure, more education and job opportunities, an extensive cultural infrastructure, and other location-specific amenities (Buch et al., 2014; Gans, 2017).

In contrast to this trend, unemployment rates in Germany are higher in the cities compared to its surroundings. The unemployment rate is one of the most important economic indicators. Unemployment has far-reaching indirect effects on the respective region: It favours the decline of wage levels, educational activities within companies, population mobility, life and health satisfaction, intelligence, and school performance, as well as rising right-wing extremism (Grözinger, 2009). The persistence of spatial disparities in unemployment in an economy is also shown by Elhorst (2003). He points out that regional unemployment is influenced by labour supply (affected by changes in the labour force, such as migration and commuting), labour demand, and wage-setting factors. Kosfeld and Dreger (2006) conduct a spatial analysis of the German regional labour market, showing that strong spatial dependencies can distort the relationship between employment and unemployment. Also Patuelli et al. (2011) include spatial linkages to effectively predict regional economic variables and to uncover spatial patterns. Particularly, there are strong relationships of dependence between cities and their surrounding areas. Identifying the cities as job magnets and finding high unemployment rates at the same

time seems contradictory. According to Grözinger (2018), this phenomenon is a 'false' effect and can be explained by the common definition of unemployment. Traditional unemployment rates are defined by the International Labour Organization (ILO) as the number of unemployed persons counted at their place of residence divided by the total number of persons in the labour force who are resident in the target area. This definition includes only the place of residence as a focal point for calculating these rates. In contrast to traditional unemployment rates, an alternative definition using the workplace as a focal point enables other insightful interpretation possibilities. Following Grözinger (2018), this alternative definition puts the resident unemployed of an area in relation to the labour force of the same area counted at the workplace. The alternative unemployment rate include commuters at their workplace and thus reflect the difference in the supply of jobs. This definition provides valuable information on missing workplaces in regional areas and support policy decisions in urban planning. Thereby, policymakers can identify regions where it might be useful to promote the settlement of companies to lower their unemployment rate and shorten commuter movements. For cities, lower alternative unemployment rates are assumed compared to the traditional definition. Low alternative unemployment rates contribute to the attractiveness of cities and the moving and commuting behaviour towards urban areas. Grözinger (2018) investigates this difference, among others, for regional areas in the German federal states Bavaria and Schleswig-Holstein. Furthermore, the comparison of both rates also provides valuable information on commuting behaviour in regional areas.

For analysing unemployment rates in the context of commuter behaviour, we look at the regional level of Functional Urban Areas (FUAs). For member countries of the Organisation for Economic Co-operation and Development (OECD), FUAs have been created as harmonised geometries describing urban areas (Dijkstra and Poelman, 2011). These regional areas are composed of city cores and their commuting zones. In this application, we use the FUAs in particular to include commuters and commuter areas to a greater extent. Hence, we are interested in considering only the city core and commuter zone separately, which is a spatial level underneath the FUA. We refer to our regional target level in the following as the FUA sublevel. This spatial level is particularly suitable for comparing the two unemployment rates described above, which differ in the spatial reference of the working population. Since this regional level is available for all OECD countries, our comparison of unemployment rates is transferable to other OECD countries and does not represent a purely German phenomenon. Furthermore, due to data availability, we only consider Germany and particularly the federal state of North Rhine-Westphalia (NRW) which is the federal state with the highest number of commuters in Germany (Bundesagentur für Arbeit, 2022b).

To estimate unemployment rates, our primary data source is the European Union Labour Force Survey (LFS). The LFS enables the estimation of both unemployment rates. The survey is designed on the governmental regions level, which is a higher regional level than the FUA sublevel (Eurostat, 2019b). According to the Nomenclature of Territorial Units for Statistics (NUTS) of the European Union, the German governmental region correspond to the NUTS 2-level and the districts to the NUTS 3-level. The FUA sublevel can be composed from the NUTS 3-level. Estimates on the spatial fine FUA sublevel that are only based on survey data (direct

estimates) are likely to have large variances due to relatively small sample sizes. To increase the accuracy of the direct estimates on lower spatial levels, small area estimation (SAE) methods can be used (see e.g., Rao and Molina, 2015; Tzavidis et al., 2018). SAE methods generally combine survey data with other data sources. For example, Costa et al. (2006), Pereira et al. (2011), and Martini and Loriga (2017) estimate unemployment rates using SAE methods by using administrative data as auxiliary information. Molina and Strzalkowska-Kominiak (2020) discuss different types of SAE estimators to calculate the percentage of people in the labour force for Swiss communes out of the LFS. They use administrative data that are provided at unit-level as auxiliary information. Similarly, Marino et al. (2019) propose semi-parametric empirical best prediction for unemployment rates that requires unit-level information. For many research questions, appropriate register or administrative data is not available. In particular, unit-level data is strictly protected. Furthermore, aggregated data is often not available at spatial finer resolutions, so that information at the target level is missing. One possibility is to use alternative data sources as covariates. Toole et al. (2015) and Steele et al. (2017) propose the usage of passively collected mobile phone data as auxiliary information, as they have a finer spatial resolution, high timeliness, and are available in real time. Basically, mobile network data can serve as a basis for producing statistics with a very high level of spatial, temporal and population coverage. For example, Steele et al. (2017) use Call Detail Records (CDRs) from the mobile network and remote sensing data for estimating poverty indices in developing countries. Toole et al. (2015) estimate changes in unemployment rates after shocks in the economy in case of mass layoffs at a plant by using mobile phone data. Moreover, Marchetti et al. (2015) have investigated solutions for a broad range of applications in using new digital data. They suggest three ways to use new digital data together with SAE techniques and show the potential of these data sources to mirror aspects of well-being and other socio-economic phenomena.

Our analyses are based on dynamic mobile network data, which is more widely available and has more information content than mobile phone data. This data source validly reflects actual commuting behaviour as well as time of day and residential population, which is important for providing auxiliary information. Since commuters and daytime population affect unemployment rates, the usefulness of these covariates for estimating the traditional and alternative unemployment rates becomes apparent. Our application combines mobile network data with data from the LFS to improve the estimation of both unemployment rates on the FUA sublevel. The aim is to compare both definitions of unemployment rates at the level of interest, thus highlighting the influence of commuters. As sample sizes are small at the FUA sublevel SAE methods are needed. From a methodological perspective, we consider the Fay-Herriot (FH) model (Fay and Herriot, 1979) using mobile network data as auxiliary information. The inverse sine transformation of the dependent variable is used frequently in literature to estimate proportions when applying the FH model (Casas-Cordero et al., 2016; Burgard et al., 2016; Schmid et al., 2017). The transformation offers the advantage of stabilization of the sampling variances and helps to approximate better the normality assumptions of the model. Casas-Cordero et al. (2016), Burgard et al. (2016), and Schmid et al. (2017) apply a naive back-transformation to obtain FH estimates and their confidence intervals on the original scale. In contrast, we use a bias corrected back-transformation following Sugasawa and Kubokawa (2017) while using as

well the inverse sine transformation. To measure the uncertainty of these specific FH estimates, we propose a parametric bootstrap procedure orientated on González-Manteiga et al. (2008) to receive not only confidence intervals but also estimates for the mean squared error (MSE). The methodology is validated with official rates based on the Urban Audit. In a model-based simulation study, we show the benefit of a bias corrected back-transformation compared to a naive one.

The paper is structured as follows: Section 5.2 defines both types of unemployment rates and explains how they deal differently with commuters. Subsequently, this section introduces the data sources for constructing these indicators. Section 5.3 describes the statistical methodology. The SAE methods and the corresponding MSE estimation is applied in Section 5.4 to estimate both unemployment rates for the German federal state NRW on FUA sublevel. Section 5.5 investigates the methodology on German data for estimating the traditional unemployment rates and compares the results with official data. Furthermore, in Section 5.6, we conduct a model-based simulation study to assess the quality of the proposed estimator. Section 5.7 discusses further research potential.

5.2 Data sources and definitions for regional unemployment rates

In this section, we first introduce the two definitions of unemployment rates each dealing differently with commuters as well as our regional target level: the FUA sublevel (Section 5.2.1). Subsequently, our two data sources are described: the LFS survey data (Section 5.2.2) and mobile network data (Section 5.2.3).

5.2.1 Traditional and alternative definition of unemployment rates

The unemployment rate according to the definition of the ILO provides an international comparable indicator (ILO, 2018). Following the ILO-definition, the traditional unemployment rate $\theta_{\text{UR}_1,i}$ for regional area i is given by

$$\theta_{\text{UR}_1,i} = \frac{\text{N}_{i,\text{unempl. (residence)}}}{\text{N}_{i,\text{unempl. (residence)}} + \text{N}_{i,\text{empl. (residence)}}}. \tag{5.1}$$

This unemployment rate is defined by the number of unemployed persons living in area i ($\text{N}_{i,\text{unempl. (residence)}}$) divided by the labour force of area i. The labour force is composed of the number of unemployed and employed persons living in area i ($\text{N}_{i,\text{unempl. (residence)}} + \text{N}_{i,\text{empl. (residence)}}$). For traditional unemployment rates, the focal point for counting employed and unemployed persons is their place of residence, where persons aged 15 to 74 are considered in the ILO-definition (ILO, 2018; Eurostat, 2018a). Please note that for reasons of comparability with German official statistics, we use the age range of 15-64 years throughout the analysis. In contrast to the traditional definition, the second definition proposed by Grözinger (2018) uses the workplace as a focal point and thus counts employed persons at the area i where their workplace is located. Since unemployed persons have no place of work, they count at area i

where they live. The definition changes to

$$\theta_{\mathrm{UR}_2,i} = \frac{\mathrm{N}_{i,\text{unempl. (residence)}}}{\mathrm{N}_{i,\text{unempl. (residence)}} + \mathrm{N}_{i,\text{empl. (workplace)}}}. \tag{5.2}$$

We refer to $\theta_{\mathrm{UR}_2,i}$ as alternative unemployment rate for area i. It is composed by the number of unemployed persons ($\mathrm{N}_{i,\text{unempl. (residence)}}$) divided by the labour force aged 15 to 64 ($\mathrm{N}_{i,\text{unempl. (residence)}} + \mathrm{N}_{i,\text{empl. (workplace)}}$). Comparing alternative unemployment rates to traditional ones, the denominator changes as employed persons count in the area i where they work. Overall, both unemployment rates treat commuters differently. If commuting is not exactly balanced, the two unemployment rates differ, and this difference reveals the influence of commuters. If the traditional unemployment rate in area i ($\theta_{\mathrm{UR}_1,i}$) is higher than the alternative one ($\theta_{\mathrm{UR}_2,i}$), there is a stronger commuter movement from other areas to area i than the other way around which is assumed for larger cities.

We focus on the alternative definition of unemployment rates as defined in Equation 5.2. However, there are, other alternative definitions such as those of the Federal Labour Office in Germany (Bundesagentur für Arbeit, 2022a) or the U.S. Bureau of Labor Statistics (U.S. Bureau of Labor Statistics, 2021), which take into account a more socio-political perspective and the relative underutilisation of the labour supply. In contrast to the alternative definition according to Grözinger (2018) used here, the labour force remains the same as in the traditional unemployment rate, while the numerator changes.

In this study, the geographical target level for investigating unemployment rates is the FUA sublevel which is particularly suitable to illustrate the difference in both definitions of unemployment rates caused by commuter flows. To the best of our knowledge, the FUA sublevel is the only OECD harmonised geometry that allows a distinction between city cores and their commuter zones. City cores are urban centres with at least 50 000 inhabitants. The commuting zone contains the surrounding travel-to-work areas of the city core where at least 15% of their employed residents are working in the respective city core (Eurostat, 2018b). Please note that the FUA sublevel as well as the FUA do not cover the whole territory of a country. Germany has in total 208 units, which are relevant for determining FUAs. These are composed of 125 city cores and 83 commuting zones. Since some commuting zones can be assigned to several city cores, there are fewer commuting zones than city cores.

5.2.2 Labour Force Survey

The LFS (Eurostat, 2019b) enables the estimation of the traditional and alternative unemployment rates introduced in Section 5.2.1. It is a household survey conducted in 35 countries including all 27 EU member states and the United Kingdom, which provides information about the labour market participation. In Germany, the LFS is part of the German Microcensus, which is a one-percent sample of the population and collected annually. All inhabitants who have their main or secondary residence in Germany and live in private or collective households are included. The sampling design corresponds to a stratified single-stage cluster sample, where neighbouring buildings are sampled and all households and persons within this cluster are surveyed. The sample districts are stratified according to region and size of the buildings

(Eurostat, 2019c). In the used LFS data, regional disaggregation is carried out using the EU-harmonised NUTS classification (Eurostat, 2018c). In Germany, the NUTS 1-level corresponds to the 16 federal states, the NUTS 2-level to 38 governmental regions, and the NUTS 3-level to the 401 administrative districts (European Parliament and Council, 2003). Traditional unemployment rates using LFS data are published on the 38 governmental regions level (NUTS 2-level). However, our target level is the smaller FUA sublevel which can be composed from the NUTS 3-level in Germany. As all LFS observations contain information about the NUTS 3-level and even finer, we can use the individual information of the LFS participants to match a) the place of residence and b) the place of work to the corresponding FUA sublevel.

In addition to the FUA sublevel, there are other possible spatial levels that are suitable to examine unemployment rates. The so-called Labour Market Areas (LMAs) are functional spatial areas that capture regional labour market structures based on commuting flows (Franconi et al., 2017). In principle, both territorial structures pursue the same goal. Nevertheless, there are practical reasons and advantages to prefer the FUA sublevel in the context of this work: First, the LMAs are compiled from commuter statistics using a specially developed algorithm. In contrast, FUAs are based on territorial structure, are already harmonised, and comparable across countries. Second, the separation of the city cores and commuter zones is an advantage of FUAs versus LMAs, which is fundamental for our analysis. Third, Germany provides indicators for the Urban Audit, which is an official statistic and publishes labour market indicators, including traditional unemployment rates, at the level of the entire FUA (one level above our target level) which we can use for external validation. All in all, the FUAs are more suitable for our analyses than the LMAs, since they fit better to the research question and are easier to handle.

In this work, we consider the year 2016 with an overall sample size of 369 986 observations in the LFS. Since the FUA sublevel does not cover the whole territory, the sample size decreases to 271 587 observations. Due to known gender differences in employment, the following analyses are carried out separately by sex. Men work more often full-time, while the proportion of women employed part-time has increased in recent years, so that overall fewer women than men are unemployed (Klammer and Menke, 2020; Statistisches Bundesamt, 2021). Due to the still existing classical gender role model, women commute fewer and shorter distances than men. These differences in behaviour justify why it is meaningful to examine unemployment separately by sex (Augustijn, 2018). Table 5.1 represents the sample sizes in the LFS based on the published NUTS 2-level and on the FUA sublevel by sex. It can be seen that the sample sizes are smaller in case of the FUA sublevel. On average, the sample sizes decrease by a factor of 7.3. Since the LFS was designed to produce reliable estimates on NUTS 2-level, the challenge of this work is to estimate reliable unemployment rates on the smaller FUA sublevel. Even if the sample sizes for FUA sublevel appear to be rather high, with a median of 368 and 421 for men and women, respectively, results in Section 5.4 show that the coefficient of variation (CV) of the direct estimates often exceeds the threshold of 20% which specifies reliable estimates at Eurostat (Eurostat, 2019a). SAE methods are discussed to obtain more reliable model-based estimates on the FUA sublevel. Since SAE methods take advantage of auxiliary variables from other data sources the auxiliary information used here is described in

Table 5.1: Distribution of sample sizes in the LFS on NUTS 2-level and FUA sublevel in Germany by sex.

	Sex	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
NUTS 2-level	Female	1 162	2 916	4 104	4 623	5 521	10 684
	Male	1 318	3 304	4 565	5 114	6 108	11 675
FUA sublevel	Female	100	216	368	635	646	7 973
	Male	97	244	421	702	749	8 559

more detail in the next Section 5.2.3.

5.2.3 Mobile network data

To estimate unemployment rates on FUA sublevel using SAE methods, we take advantage of suitable auxiliary information. Many SAE applications are based on register data as a second data source. These data sources are not timely or are aggregated to higher (regional) levels. Alternative data sources have the potential to overcome these disadvantages. For example, Toole et al. (2015) or Steele et al. (2017) have used mobile phone data for SAE. Mobile network data are explored to estimate daytime population, commuter patterns or tourism behaviour (see e.g., De Meersman et al., 2016; Galiana et al., 2018). Mobile network data represent mobile activities or signals from the mobile network of the respective mobile network operator. A mobile activity is defined as an event caused by a length of stay in a specific geometry without movement (also known as dwell time). Signalling data are produced automatically, regularly and only register the location of the cell tower to which a mobile device is connected at a specific time. Therefore, they are collected as a by-product and tend to be less costly compared to official survey data. The major advantage of these data sources are their real-time availability, high temporal actuality, nationwide availability, and their finer spatial resolution. Mobile activities can be obtained at the spatial resolution of cities, communities or grid cells, so that a simple assignment to other spatial levels such as the FUA sublevel is possible. This spatial flexibility and high resolution are not feasible with register or administrative data. In many countries, like Germany, register data are strictly protected and thus not available at high resolution or on specific regional levels. In addition, mobile network data are dynamic, so that the movement of activities can be observed over the course of the day as well as daily, during a week or a month. Previous analyses in Germany have shown that mobile network data correlate strong with register-based census data like population figures and with the population mobility, more precisely commuter movements (Hadam, 2018, 2021). This is, among other things, due to the high penetration rate of mobile devices in the German population (Statistisches Bundesamt, 2022). Accordingly, mobile network data provide a reliable picture of the real physical locations of the German daytime and night-time population or with other words the resident and working population compared to official statistics with a fixed reporting date. Since our aim is to estimate an alternative unemployment rate accounting for commuters mobile network data reflecting resident and working population are especially suitable auxiliary information (cf. Hadam (2021)).

In Germany, there are three mobile network operators: Deutsche Telekom, Vodafone, and

Telefónica Deutschland, with a respective market share of one-third each. The data records available to Destatis and used for this work contain mobile activities of Deutsche Telekom customers. In compliance with data protection rules, the mobile activities are anonymised and aggregated. Regionally fluctuating market shares of each mobile network operator are adjusted regionally as part of the extrapolation procedure at the respective operator. Thus, the estimated local market shares of each operator are used as weights to adjust the mobile network data. The data records include contract, prepaid, and further customers. In addition, mobile network data contain information on socio-demographic characteristics of mobile device users, such as age group, sex, and nationality of the SIM card owner. However, the characteristics are only available for contract customers. Furthermore, the following assumptions were made in the data provider's data generation process: Since the number of mobile activities depends on the dwell time of mobile devices, long mobile device activities are counted and included in the data record according to the length of the dwell time, while short mobile activities are not considered. The dwell time in the data record available is two hours to filter out short mobile device activities (for example, quick movements between the grid cells). Finally, only values based on a minimum number of 30 activities per geometry were provided due to data protection reasons.

Our aim is to analyse the effect of commuters on the two proposed unemployment rates. Since we use a model-based method, suitable covariates are crucial. We only use mobile network data for this purpose and no further covariates. As we will show in Section 4.1, our models with only mobile network covariates lead to high coefficients of determination, so mobile network data are sufficient as SAE covariates in our case.

We define from the mobile network data 27 auxiliary variables. Between 7 to 16 auxiliary variables are chosen by model selection procedure (cf. Section 5.4.1). The data contains mobile activities for a statistical week that consists of 24-hour days. These were selected from the months April, May, and September in 2017 without school or public holidays to avoid distortions in the representation of commuters. The mobile activities comprise the average activities on the selected weekdays. The weekdays are categorised according to five types of days, with the days from Tuesday to Thursday being grouped together. Since the counted activities of mobile devices alone are not meaningful enough, further covariates are constructed from the available mobile network data at the FUA sublevel. The aim in creating the covariate is to highlight the differences between the daily and resident population and thus the commuters themselves. This is particularly reflected in the changes in the intensities of mobile activities. Based on this, covariates are calculated in the form of ratios, shares, and change values which reflect exactly these differences. Since it is assumed that the unemployed persons are more likely to stay at home during the day and the employed are more likely to stay at the place of work, the rate and change of activities in the morning and evening hours are calculated. This means, that the change from place of work to the place of residence and vice versa is modelled. This includes the change in mobile activities of working hours and hours spent at home as well as the change in activities of potential commuters. In addition, the change in activities during the day is calculated and the differences in core times or peaks in mobile activities are determined. The core times are based on the usual working times in Germany (7 am

to 4 pm). Furthermore, differences in mobile network activities among to socio-demographic characteristics such as age, nationalities (summarised by continent), and sex can also be considered. These characteristics also have an influence on commuting behaviour. An overview of the selected mobile network covariates can be found in the supplementary material in Table C.1.

5.3 Small area method

In this section, the statistical methodology for estimating unemployment rates on FUA sublevel is described. As the LFS is designed for higher regional levels, a model-based approach enriched by auxiliary variables from mobile network data is used. We use the FH model (Fay and Herriot, 1979), an area-level model that links direct estimates to area level covariates. The FH model is especially useful in countries with strict data protection requirements like Germany, as the auxiliary variables and the direct estimates only need to be available on an aggregated level. As in Casas-Cordero et al. (2016), Burgard et al. (2016), and Schmid et al. (2017) we use the inverse sine transformation on the dependent variable to estimate proportions using area-level models. Following Sugasawa and Kubokawa (2017), we derive the inverse sine transformed FH model including a bias correction for the back-transformation. A parametric bootstrap, which incorporates the bias correction, is proposed.

5.3.1 Fay-Herriot estimates

In the following, we assume a finite population of size N, which is divided into d areas. The present sample consists of areas with different sample sizes $n_1, ..., n_d$ drawn by a complex design from the population. To refer to the actual area, we use the subscript i. The population size and sample size of this area is indicated with N_i and n_i, respectively. The FH model is a special case of a linear mixed model. Please note that typical linear mixed models use two indices to identify individuals within specific groups while the FH model has only one index. The FH model links a vector with p area-specific covariates $\mathbf{x}_i$ to the direct estimate ($\hat{\theta}_i^{\text{direct}}$) using an area-specific random effect u_i for each area $i \in 1, ..., d$:

$$\hat{\theta}_i^{\text{direct}} = \mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i, \quad u_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2) \quad \text{and} \quad e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_{e_i}^2).$$

The model assumes that the random effects u_i are identically independently normally distributed and the sampling errors e_i are independently normally distributed. $\hat{\theta}_i^{\text{direct}}$ is the direct estimate for the unemployment rates for a certain area i and $\hat{\sigma}_{e_i}^2$ its variance estimate that are estimated with the **survey** package from R (Lumley, 2004; R Core Team, 2022) considering the sampling design of the LFS and the survey weights. The regression parameters $\hat{\boldsymbol{\beta}}$ can be estimated as best linear unbiased estimator of $\boldsymbol{\beta}$ and the random effect $\hat{u}_i$ as empirical best linear unbiased predictor of u_i (Rao and Molina, 2015). For the estimation of the variance of the random effects σ_u^2, several approaches are available: The FH method of moments, the maximum likelihood method (ML), and the restricted maximum likelihood method (REML) among others (Rao and Molina, 2015). For our analysis, we use the REML method.

Through this combination, we obtain the resulting FH estimator, which is an empirical best linear unbiased predictor of θ_i. It is as a weighted combination of the direct estimator $\hat{\theta}_i^{\text{direct}}$ and the synthetic estimator $\mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ for each area i:

$$\begin{aligned}\hat{\theta}_i^{\text{FH}} &= \mathbf{x}_i^T\hat{\boldsymbol{\beta}} + \hat{u}_i \\ &= \hat{\gamma}_i\hat{\theta}_i^{\text{direct}} + (1-\hat{\gamma}_i)\mathbf{x}_i^T\hat{\boldsymbol{\beta}},\end{aligned} \tag{5.3}$$

where the shrinkage factor $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\sigma_{e_i}^2}$ defines the weight on both parts for each area i. Whenever the variance of the sampling errors is relatively small for a specific area i, more weight is assigned on its direct estimator.

5.3.2 Back-transformed Fay-Herriot estimates

As unemployment rates are a percentage, we transform the dependent variable to profit from the variance stabilization of the sampling variance. Thus, we use the inverse sine transformation $h(x) = \sin^{-1}(\sqrt{x})$ as in Casas-Cordero et al. (2016), Burgard et al. (2016), and Schmid et al. (2017). Note that Schmid et al. (2017) compared in a design-based simulation study the inverse sine transformation with alternative modelling options, for instance an estimator based on a normal-logistic distribution. Both estimators lead to very similar results regarding MSE and bias. Raghunathan et al. (2007) defends the choice of the inverse sine transformation for estimating cancer risk factors rates against generalized linear models with their higher complex design features and computational tasks. While they all use a naive back-transformation $h^{-1}(x) = \sin^2(x)$, we transform the FH estimator back to the original level with consideration to the back-transformation bias. Burgard et al. (2016) mentioned the methodology for a bias corrected back-transformation. We derive the back-transformation following Sugasawa and Kubokawa (2017), who introduce the FH model for general transformations on the dependent variable. Following Jiang et al. (2001), we approximate the sampling variances of the transformed direct estimates by $\tilde{\sigma}_{e_i}^2 = 1/4\tilde{n}_i$, where $\tilde{n}_i$ denotes the effective sample size. The design effects and thus the effective sample size can also be estimated with the **survey** package (Lumley, 2004; R Core Team, 2022). For the model on the transformed scale, we consider the assumptions of the FH model

$$\sin^{-1}\left(\sqrt{\hat{\theta}_i^{\text{direct}}}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + u_i + e_i, \quad u_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma_u^2) \text{ and } e_i \overset{\text{ind}}{\sim} \mathcal{N}(0,\tilde{\sigma}_{e_i}^2). \tag{5.4}$$

Out of the FH model on transformed scale in Equation 5.4, $\hat{\boldsymbol{\beta}}$ and $\hat{u}_i$ can be estimated, as described in the previous Section 5.3.1. Replacing the model parameters with their estimates leads to the FH estimator on the transformed level:

$$\hat{\theta}_i^{\text{FH}^*} = \hat{\gamma}_i \sin^{-1}\left(\sqrt{\hat{\theta}_i^{\text{direct}}}\right) + (1-\hat{\gamma}_i)\mathbf{x}_i^T\hat{\boldsymbol{\beta}}.$$

However, the goal is to get the FH estimator on the original scale $\left(\hat{\theta}_i^{\text{FH, trans}}\right)$. For this reason, $\hat{\theta}_i^{\text{FH}^*}$ must be back-transformed. According to the Jensen-inequality (Jensen et al., 1906), a naive back-transformation $\left(\sin^2\left(\hat{\theta}_i^{\text{FH}^*}\right)\right)$ leads to biased results due to the non-linearity of the

transformation. To avoid this bias, the following formula using the known distribution of the FH estimator on the transformed level $\hat{\theta}_i^{\text{FH}^*} \sim \mathcal{N}\left(\hat{\theta}_i^{\text{FH}^*}, \frac{\hat{\sigma}_u^2 \tilde{\sigma}_{e_i}^2}{\hat{\sigma}_u^2 + \tilde{\sigma}_{e_i}^2}\right)$ is used

$$\begin{aligned}
\hat{\theta}_i^{\text{FH, trans}} &= E\left\{\sin^2\left(\hat{\theta}_i^{\text{FH}^*}\right)\right\} \\
&= \int_{-\infty}^{\infty} \sin^2(t) f_{\hat{\theta}_i^{\text{FH}^*}}(t) dt \\
&= \int_{-\infty}^{\infty} \sin^2(t) \frac{1}{2\pi \frac{\hat{\sigma}_u^2 \tilde{\sigma}_{e_i}^2}{\hat{\sigma}_u^2 + \tilde{\sigma}_{e_i}^2}} \exp\left(-\frac{\left(t - \hat{\theta}_i^{\text{FH}^*}\right)^2}{2\frac{\hat{\sigma}_u^2 \sigma_{e_i}^2}{\tilde{\sigma}_u^2 + \tilde{\sigma}_{e_i}^2}}\right) dt, \qquad (5.5)
\end{aligned}$$

where $\hat{\theta}_i^{\text{FH, trans}}$ denotes the transformed FH estimator. To solve this integral, numerical integration techniques are applied. In Section 5.6, the proposed bias corrected FH estimator ($\hat{\theta}_i^{\text{FH, trans}}$) is evaluated in a close to reality model-based simulation study.

5.3.3 Uncertainty estimation

As a measurement of uncertainty for $\hat{\theta}_i^{\text{FH, trans}}$, a parametric bootstrap MSE as well as parametric bootstrap confidence intervals are constructed. When using a FH model without transformations or with a log transformation, analytical solutions to estimate the MSE are known (Prasad and Rao, 1990; Datta and Lahiri, 2000; Slud and Maiti, 2006). Up to our knowledge, no analytical solution is available in the case of the inverse sine transformation. Bootstrap methods are very promising to estimate the MSE. Casas-Cordero et al. (2016) construct confidence intervals using a parametric bootstrap procedure, in which confidence interval limits are built on the transformed scale with subsequent naive back-transformation for each bootstrap replication. In contrast to this methodology, our goal is to construct confidence intervals and a MSE for FH estimates from a model using the inverse sine transformation. Another difference is that, instead of the naive back-transformed FH estimates, the bias corrected back-transformed FH estimates are included within the bootstrap procedure. Our parametric bootstrap is orientated on the bootstrap procedure of González-Manteiga et al. (2008). In the following, the steps of the used bootstrap method to construct both measurements of uncertainty are shown:

- From the model on the transformed scale (Equation 5.4), take $\tilde{\sigma}_{e_i}^2$ and estimate $\hat{\sigma}_u^2$ and $\hat{\boldsymbol{\beta}}$ using the sample data.
- For $b = 1, ..., B$
  - Generate area specific random effects $u_i^* \sim \mathcal{N}(0, \hat{\sigma}_u^2)$ and sampling errors $e_i^* \sim \mathcal{N}(0, \tilde{\sigma}_{e_i}^2)$.
  - Bootstrap samples:
    - Use u_i^* and e_i^* to construct the bootstrap sample on the transformed scale

$$\sin^{-1}\left(\sqrt{\hat{\theta}_{i,(b)}^{\text{direct}}}\right) = \mathbf{x}_i^T \boldsymbol{\beta} + u_i^* + e_i^*.$$

    - Use the bootstrap sample to estimate the FH estimator on the transformed scale $\left(\hat{\theta}_{i,(b)}^{\text{FH}^*}\right)$ as described in Section 5.3.2.

  - ∗ Determine the FH estimates on the original scale $\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}}\right)$ using (Equation 5.5) to account for the bias correction.

- Bootstrap population:
  - ∗ Use u_i^* to construct the bootstrap population on the transformed scale

$$\sin^{-1}\left(\sqrt{\hat{\theta}_{i,(b)}^{\text{direct}}}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + u_i^*.$$

  - ∗ For each bootstrap population, calculate the population mean on the original scale

$$\theta_{i,(b)}^{\text{trans}} = \sin^2\left(\mathbf{x}_i^T\boldsymbol{\beta} + u_i^*\right).$$

- Predict the MSE and the 95% confidence intervals

$$\text{MSE}(\hat{\theta}_i^{\text{FH, trans}}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}} - \theta_{i,(b)}^{\text{trans}}\right)^2 \tag{5.6}$$

$$\text{CI}(\hat{\theta}_i^{\text{FH, trans}}) = \Big[\hat{\theta}_i^{\text{FH, trans}} + q_{0.025}\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}} - \theta_{i,(b)}^{\text{trans}}\right); \hat{\theta}_i^{\text{FH, trans}} + q_{0.975}\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}} - \theta_{i,(b)}^{\text{trans}}\right)\Big], \tag{5.7}$$

where $q_{0.025}$ is the 2.5% quantile over the bootstrap replications and $q_{0.975}$ respectively the 97.5 % quantile.

The methodology presented above for constructing uncertainty measurements for the back-transformed FH estimates is also evaluated within a simulation study (cf. Section 5.6).

5.4 Alternative unemployment rates including commuters in North Rhine-Westphalia

In this section, we determine and discuss traditional and alternative unemployment rates that deal differently with commuters. For this purpose, we use the LFS data from Section 5.2.2 and the mobile network data from Section 5.2.3. Traditional and alternative unemployment rates have been introduced in Section 5.2.1. The members of the labour force are counted for the two rates at different reference points: At the place of residence (traditional unemployment rates) or at the place of work (alternative unemployment rates). In particular, they assign commuters to different small areas. When using traditional unemployment rates, the contradiction of high unemployment rates in the city cores results from the exclusion of commuting. Alternative unemployment rates are expected to exceed traditional ones in commuter zones and to be lower in city cores. We confirm this empirically. The rates are estimated separately by sex and at the target level of the FUA sublevel.

5.4.1 Model selection and validation

Four models need to be created and validated. Following Schmid et al. (2017), the Bayesian information criterion for a simple linear regression model is used for the model selection. As

Table 5.2: Measurements to validate the FH models for traditional (UR_1) and alternative unemployment rates (UR_2) separated by sex: This table shows the estimated variance of the random effects ($\hat{\sigma}_u^2$), the Shapiro-Wilks (S.-W.) p-value for level 1 and level 2 error terms as well as the modified R^2.

	Men		Women	
	UR_1	UR_2	UR_1	UR_2
$\hat{\sigma}_u^2$	0.000320	0.000361	0.000716	0.000880
S.-W. p-value: level 1	0.308668	0.495064	0.809323	0.866098
S.-W. p-value: level 2	0.695112	0.549476	0.861257	0.901708
modified R^2	0.772521	0.908642	0.632059	0.575550

dependent variable, we use the inverse sine transformed direct estimates from LFS and the auxiliary information is mobile network data (cf. Section 5.2.3). In total, 6 to 16 of 27 potential mobile network covariates are selected depending on the model. The covariates of all four models are listed in Table C.1 within the Appendix C.1. Since the models are built on the transformed scale, the coefficients have no natural interpretation in terms of expected values at the original level, but their direction is directly interpretable. The chosen covariates reflect most likely relationships between working and non-working hours and the changes in mobile activities due to commuting during the day and evening. The latter is represented less strongly in the females model, which is in line with lower commuting patterns of women. An increase of covariates that proxy possible commuter movements generally leads to a decrease of alternative unemployment rates (UR_2). The reverse is the case for traditional ones (UR_1). All models include changes from night to day activities of other nationalities, most likely tourists, which have a positive impact on regional employment. As expected, negative values have been observed for these coefficients.

To investigate the explanatory power of the models, we use the modified R^2 from Lahiri and Suntornchost (2015) and obtain values of at least 57% as shown in Table 5.2. Furthermore, we check whether meaningful results are obtained for estimating the variance of the random effects using REML estimation. As Table 5.2 shows, positive values were estimated in all cases. Thus, the potential problem of negatively estimated variances does not occur. For each FH model on the transformed scale, the assumptions on the error terms (level 1 and 2) are checked. The normality assumptions of the random effects (level 2) as well as of the residuals (level 1) - obtained from fitting the model (Equation 5.4) - are tested. The p-values of the Shapiro-Wilks test in Table 5.2 confirm that in all cases the normality assumption for both error terms cannot be rejected. Overall, all four models could be validated and are suitable for subsequent analyses.

5.4.2 Gain in accuracy

To assess the gain in the reliability of the estimators, we compare the CVs. Figure 5.1 visualises this measurement for the different methods and definitions of unemployment rates. Eurostat considers estimators with a CV below 20% to be reliable (Eurostat, 2019a). If we use direct estimation 53.7% (men; UR_1), 29.3% (women; UR_1), 53.7% (men; UR_2), and 31.7% (women;

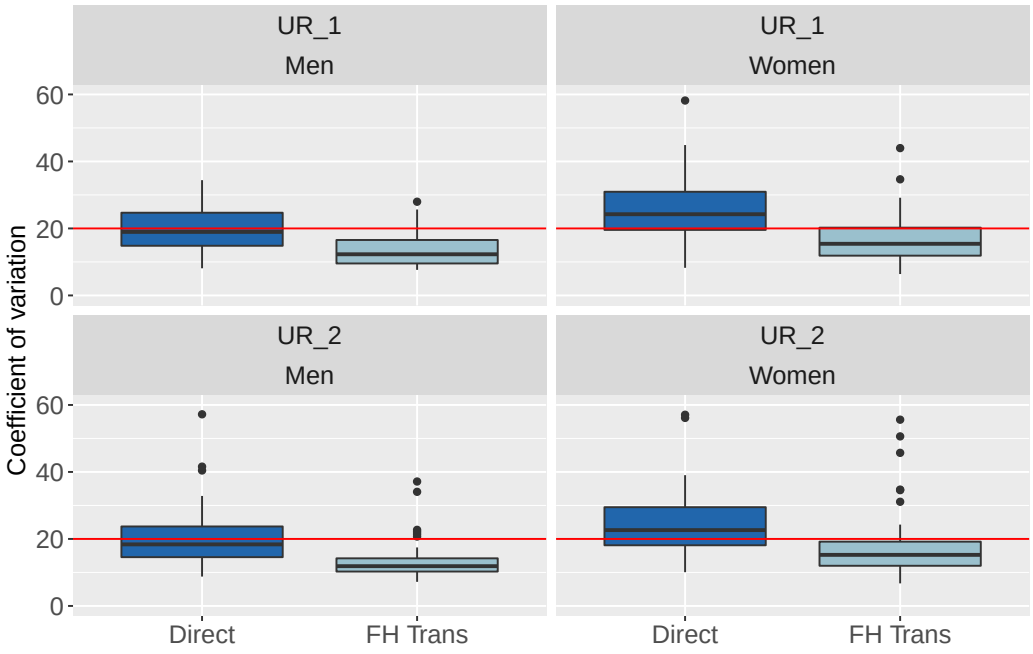


Figure 5.1: Reduction of the coefficient of variation by using the transformed FH model instead of direct estimation for estimating unemployment rates in NRW.

UR_2) of the CVs are below 20%. The use of the transformed FH model achieves a distinct increase of CVs below this threshold. As a result, 85.4% (men; UR_1), 73.2% (women; UR_1), 82.9% (men; UR_2), and 78.0% (women; UR_2) of the CVs are below 20%. This illustrates that the use of dynamic mobile network data in combination with SAE methodology is a powerful tool to increase the precision of both estimated unemployment rates for NRW on FUA sublevel. If we compare the direct estimates to the estimates from the proposed transformed FH model, both are often close to each other. For regions with smaller samples sizes like Witten and Paderborn, these values can deviate clearly from each other. Due to the higher uncertainty of the direct estimates for regions with lower sample sizes, the synthetic part within Equation 5.3 is weighted higher and bigger differences to the direct estimates appear.

5.4.3 Discussion of the estimated unemployment rates for NRW

Figure 5.2 illustrates the differences between the alternative and traditional unemployment rates. If the traditional unemployment rates are the same as alternative one, the commuter behaviour is balanced and the calculated difference would be zero. Please note, that the FUA sublevels do not cover the entire federal territory in NRW, these areas are white in Figure 5.2. The bluish colors indicate areas where the alternative unemployment rate is higher than the traditional one. Those are mainly the commuter zones in both models, i.e., the commuter flow is directed out of this area. With one exception in the female model, all commuting zones are coloured blue. This means that these areas are the place of residence of many employed people who commute from those areas to their workplace. The reddish areas, however, imply that the alternative unemployment rate is lower than the traditional unemployment rate. This is mainly the case for the city cores of the FUAs. This observation is consistent with

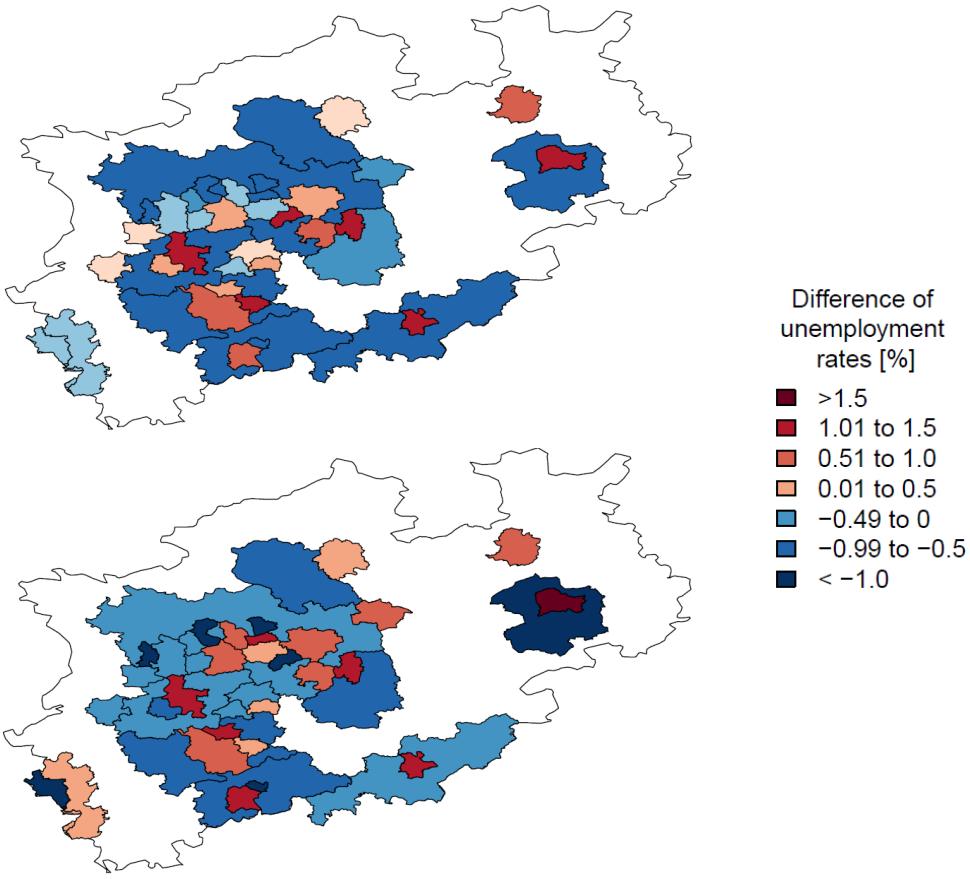


Figure 5.2: Difference of unemployment rates due to including commuters for men (above) and women (below). The spatial assignment of city names to the FUA sublevels is shown in the Appendix C.2.

Grözinger (2018) motivation for creating an alternative unemployment rate. Nevertheless, a negative value (blue colouring) was detected for a few city cores. This is the case for nine city cores simultaneously in both models. These are the city cores Recklinghausen, Bottrop, Moers, Oberhausen, Duisburg, and Mühlheim an der Ruhr. These six are located in the Ruhr region, which includes the large city cores Essen and Dortmund, to which many people commute from the Ruhr region. Furthermore, this trend was found for the two small city cores (Solingen and Sankt Augustin) and Aachen, which is located directly on the Belgian border. Since most city cores are job engines, many employed people living in the surrounding travel-to-work areas, which is their place of residence, commute into the city cores to work. In the males model, the differences are higher than in the females model, which leads to the conclusion that women are not commuting as often or as far as men (IT.NRW, 2019). Possible reasons for this could be the conservative role model of women, the spatial closeness to the family that is guaranteed by the woman (to the school/kindergarten of the children, etc.) or, for example, a work in small, nearby companies/enterprises (Bauer-Hailer, 2019).

5.5 Validity of the proposed method

In the following, we evaluate the methodology used in Section 5.4 to estimate unemployment rates at the FUA sublevel through official data. For Germany, the database Urban Audit provided by Eurostat in cooperation with Destatis and Kommunales Statistisches Informationssystem (KOSIS) is the only source for German unemployment rates at the FUA level (KOSIS-Gemeinschaft Urban Audit, 2013; Eurostat, 2017, 2019d). This official data source provides traditional unemployment rates, but no alternative unemployment rates for all German FUAs. Thus, the Urban Audit enables a comparison of traditional unemployment rates estimated by using the transformed FH estimator (Equation 5.4) with mobile network data as auxiliary information with the officially published values. As mentioned in Section 5.2.1, we have used the 15-64 age range for the definitions of unemployment rates to ensure comparability with the Urban Audit. Please note, the comparison in this section is made on the entire FUA level and not on the FUA sublevel as in the application in Section 5.4.

For the German federal state NRW, we have an extensive mobile network data record available as auxiliary information. However, we have only limited access to mobile network data and accordingly a data set with less information for the rest of the country. Thus, less covariates are available for the validation. In contrast to Section 5.4, where we use dynamic signalling data, we only have static mobile network activities of a typical Sunday evening for the whole of Germany. We focus on the time period from 8 to 11 pm of the average of eight Sundays of the months April, June, and July in 2018 without school or public holidays. For Sunday evenings, a high correlation has been identified between population figures from the 2011 census and the mobile network activities on the weekend and especially on Sunday evening (Hadam, 2018). As traditional unemployment rates are based on the place of residence, it is reasonable to assume that mobile network data of a Sunday evening is suitable as auxiliary variables. In the following, we validate the proposed transformed FH model by comparing the FH estimates with official unemployment rates of the Urban Audit. We use the SAE method and model selection as applied in Section 5.4 with the difference that a) the regional focus is now FUAs across Germany and b) we can only use mobile network data from Sunday evening. In the males model, the selected mobile network covariates explain around 47% of the variance in terms of the modified R^2 following Lahiri and Suntornchost (2015) and in the females model around 37%.

For the validation of the proposed method, Figure 5.3 shows the estimated unemployment rates using mobile network covariates (FH Trans), the direct, and the published official estimates from Urban Audit by sex. First, it can be seen that we get similar rates compared to the Urban Audit by using the transformed FH model. Comparing the direct estimator from the LFS with the FH Trans estimator, the FH Trans estimator corrects the direct estimator in such a way that the resulting value is closer to the Urban Audit. This trend is quantified in Table 5.3. It reports the distribution of the absolute difference of the females and males unemployment rates obtained by the two estimation methods for all FUAs in Germany compared to the Urban Audit. For almost all distribution values, we get a higher absolute difference for the direct estimates compared to the FH Trans estimates. Only in the males model the 25% quantile for the absolute difference is slightly higher for the FH Trans estimates. As expected, it can be noted

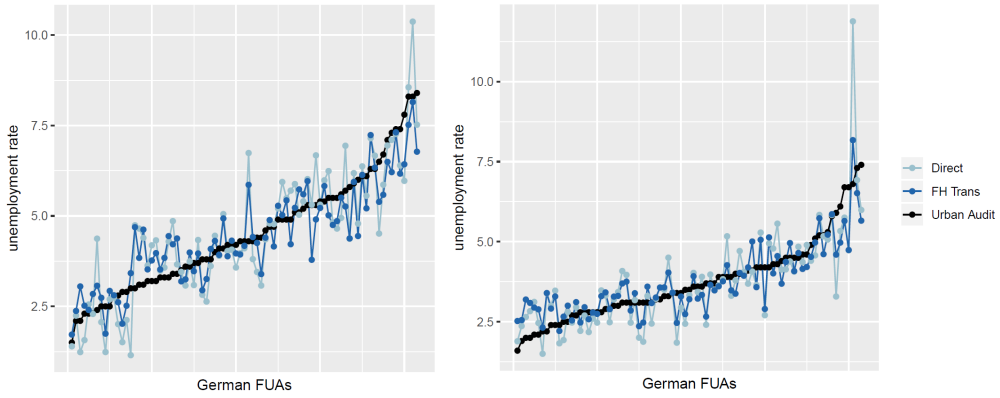


Figure 5.3: Comparison of traditional unemployment rates (UR_1) published in Urban Audit (black), estimated with the transformed FH model (dark blue) and the direct estimates from the LFS (light blue) for men (left) and women (right) for all German FUAs.

Table 5.3: Distribution of the absolute difference to the Urban Audit estimates of the females and males traditional unemployment rates over all German FUAs and in particular over FUAs with small sample sizes below 600.

Areas	Sex	Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
All	Female	Direct	0.017	0.246	0.459	0.638	0.800	5.078
		FH Trans	0.005	0.173	0.415	0.512	0.748	1.959
	Male	Direct	0.009	0.202	0.625	0.713	0.998	2.440
		FH Trans	0.008	0.221	0.428	0.573	0.824	1.690
Sample size <600	Female	Direct	0.030	0.416	0.628	0.930	1.120	5.078
		FH Trans	0.015	0.281	0.516	0.627	0.896	1.959
	Male	Direct	0.068	0.697	1.095	1.129	1.764	2.073
		FH Trans	0.038	0.373	0.676	0.704	1.027	1.690

that for FUAs with sample size under 600 estimated unemployment rates of both estimation methods show higher values for the absolute difference.

5.6 Model-based simulation

In the previous two sections, we use the proposed transformed FH model to estimate alternative unemployment rates and subsequently evaluate the suggested methodology with official statistics obtained from Urban Audit. This model-based simulation study is used to investigate how much we benefit from the more complicated transformed FH model with a bias corrected back-transformation compared to the naive back-transformation. According to the Jensen-inequality (cf. Section 5.3.2), the naive back-transformation is biased under the inverse sine transformation. Furthermore, we want to show, that the proposed MSE and confidence intervals lead to reasonable results. We investigate these aims in a close to reality environment. The input values of the model-based setting are based on the real data.

The simulation study is implemented with $R = 1\,000$ Monte-Carlo replications. Within each replication, we generate the covariates ($\mathbf{x}_i$) initially from a lognormal distribution with parameters $(-0.5, 0.04)$. The number of areas is fixed to the number of the FUA sublevels

Table 5.4: Distribution of important parameters in the simulation setting: The sampling error variation σ_{e_i} and the resulting shrinkage factor γ_i coincide with the male model for Germany on FUA sublevel. The direct estimates $\left(\hat{\theta}_i^{\text{direct}}\right)$ of the simulation study are close to the values for the FUA sublevel.

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
σ_{e_i}		0.0063	0.0202	0.0275	0.0288	0.0366	0.0785
γ_i		0.1199	0.3848	0.5265	0.5355	0.6730	0.9548
$\hat{\theta}_i^{\text{direct}}$	sim.	0.0000	0.0340	0.0495	0.0538	0.0688	0.2826
	FUA sublevel	0.0054	0.0328	0.0484	0.0508	0.0647	0.1134

in Germany ($d = 208$). We draw the random effect and the sampling errors from normal distributions: $u_i \sim \mathcal{N}(0, \sigma_u^2)$ and $e_i \sim \mathcal{N}(0, \sigma_{e_i}^2)$. According to the males model for Germany on FUA sublevel, $\sigma_u \approx 0.029$ is defined analogously. In addition, we adopt the variation of the sampling errors σ_{e_i} and keep them constant over the replications. The regression coefficients are set to $\beta_0 = 0.01$ and $\beta_1 = 0.35$. As data generating process, we consider $\hat{\theta}_i^{\text{direct}} = \sin^2\left(\beta_0 + \mathbf{x}_i^T\beta_1 + u_i + e_i\right)$ to get synthetic direct estimates. The true small area means are $\bar{y}_i = \sin^2\left(\beta_0 + \mathbf{x}_i^T\beta_1 + u_i\right)$. Table 5.4 shows the distribution of the variation of the sampling errors and the resulting shrinkage factor as well as the distribution of the direct estimates for the simulation (over all replications) and the actual direct estimated unemployment rates for males in Germany. The distributions are close to each other.

For each replication, we estimate small area means from the transformed FH model: With respect to the back-transformation bias $\left(\hat{\theta}_i^{\text{FH,trans}}\right.$, cf. Equation 5.5$\left.\right)$ and with naive back-transformation $\left(\hat{\theta}_i^{\text{FH,naive}}\right)$. To assess the quality of the estimates, we obtain for $R = 1\,000$ Monte Carlo replications the absolute Bias (aB) and the root mean squared error (RMSE) of the estimates, defined as

$$\text{aB}_i = \left| \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}_i^{\text{FH},(r)} - \bar{y}_i^{(r)} \right) \right| * 100$$

$$\text{and RMSE}_i = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}_i^{\text{FH},(r)} - \bar{y}_i^{(r)} \right)^2} * 100,$$

where $\hat{\theta}_i^{\text{FH},(r)}$ is the estimated respective FH value and $\bar{y}_i^{(r)}$ the true value within replication r. Figure 5.4 shows the reduction of aB. For instance, the median of the aB using a naive back-transformation is 1.86 times higher than with a bias corrected back-transformation. At the same time, we observe nearly the same RMSE (cf. Figure 5.4) when we use a bias corrected back-transformation instead of a naive back-transformation. In summary, there is a clear reduction in bias at the cost of a slightly higher RMSE.

We next investigate the properties of the proposed MSE and the confidence intervals. Please note that we compare the bootstrap estimated RMSE (Equation 5.6) to the empirical RMSE, which we treat as the true one. For calculating these uncertainty measurements, we use 1 000 bootstrap replications within each Monte Carlo run. As quality measurements, we calculate the relative bias of the uncertainty estimation (rB RMSE) and the relative RMSE of the uncertainty

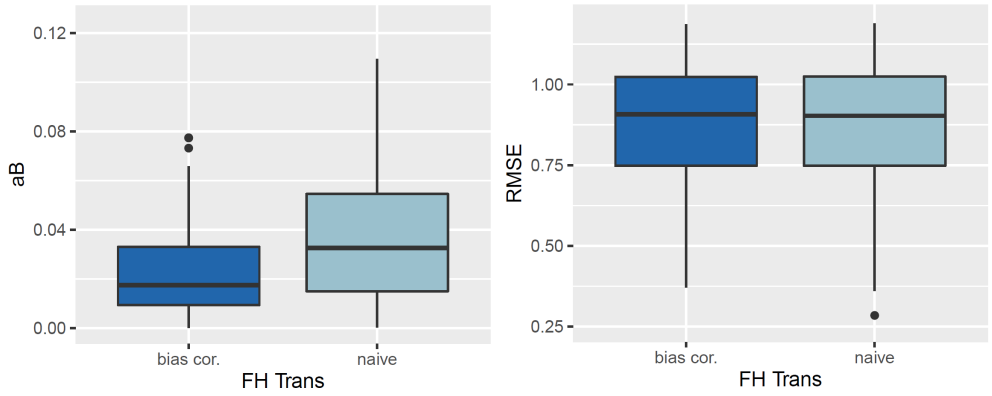


Figure 5.4: Distribution of the aB and the RMSE for the transformed FH estimator with bias corrected and naive back-transformation.

Table 5.5: Distribution of the quality measurements for the estimated RMSE and the corresponding confidence intervals using the bootstrap procedure as described in Section 5.3.3

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
rB RMSE	-9.34	-2.12	-0.62	-0.55	1.11	7.13
rRMSE RMSE	17.04	18.03	18.53	18.74	18.99	44.97
Coverage	86.70	93.90	94.40	94.34	94.90	96.00

estimation (rRMSE RMSE). They are defined as

$$\text{rB RMSE}_i = \left(\frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R} \text{MSE}_{\text{est},i}^{(r)}} - \text{RMSE}_{\text{true},i}}{\text{RMSE}_{\text{true},i}} \right) * 100$$

$$\text{and rRMSE RMSE}_i = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R} \left(\text{RMSE}_{\text{est},i}^{(r)} - \text{RMSE}_{\text{true},i} \right)^2}}{\text{RMSE}_{\text{true},i}} * 100,$$

where $\text{RMSE}_{\text{est},i}^{(r)}$ is the estimated RMSE out of the bootstrap procedure (cf. Section 5.3.3) for each Monte Carlo replication r and $\text{RMSE}_{\text{true},i}$ is the empirical RMSE over the Monte Carlo replications. The relative bias is close to zero as Table 5.5 shows. On average, we get an underestimation of 0.55% over all areas. The interquartile range goes from -2.12% to 1.11%. In addition, the relative RMSE of the estimated RMSE is important to assess its quality. We get a mean relative RMSE of 18.74% for the estimated RMSE. The low bias and the RMSE show that the proposed MSE estimator yields good results. In addition to the MSE, we can also get bootstrap confidence intervals (cf. Section 5.3.3). The coverage is defined as the proportion of the time that the estimated confidence interval contains the true value. For the proposed confidence intervals (Equation 5.7), we get in mean a coverage of 94.34%. We can recognize a slight underestimation of the coverage, but the values are close to the target value of 95%. These three measures show that the proposed bootstrap-estimated MSE works.

Overall, our close to reality simulation study shows the reduction of bias while using the transformed FH estimator with bias corrected back-transformation instead of a naive back-transformation. Furthermore it demonstrate the good performance of the newly proposed MSE

estimator and confidence intervals for the transformed FH estimator with bias corrected back-transformation.

5.7 Concluding remarks

The traditional unemployment rate is based on the place of residence of the labour force. Due to the high level of commuting, this may give a distorted impression of regional labour markets. For Germany, traditional unemployment rates show higher rates in city cores compared to its surroundings. For analysing unemployment rates in the context of commuter behaviour, the regional target area are city cores and their commuting zones, which can be extracted from FUAs. In this work, we estimate an alternative unemployment rate, where the focal point of the labour force is their workplace. It adjusts the traditional definition by including commuters. Since the LFS is not designed to produce indicators on smaller areas than NUTS 2-level, a FH approach is used to estimate alternative and traditional unemployment rates on the FUA sublevel. From a methodological point of view we use a bias corrected back-transformed FH estimator and propose a MSE estimator to measure its uncertainty. As the FH approach relies on a model-based method, suitable covariates are required. We select covariates constructed from dynamic mobile network data and validate the selected models. The benefit of dynamic mobile network data is that they represent the changes of the counted aggregated mobile devices during the day and in space. This information can be used to derive the commuting behaviour of the population. The resulting differences between the traditional and the alternative unemployment rates show that the rates in city cores are mainly lower than officially indicated. The assumption that unemployment rates in city cores are lower can be confirmed and thus contributes to the explanation why so many people move to city cores due to more job opportunities. Furthermore, the alternative definition of the unemployment rate removes the static picture of the population, especially of the labour force. The labour force does not necessarily live in the same place where they work. This dynamic cannot be achieved with traditional survey methods and with traditional data. However, exactly this knowledge is necessary to make better decisions regarding urban planning. Moreover, these alternative rates provide potential employers with additional information about the current regional labour market and on missing workplaces. This will help to identify regions for which it might be useful to promote business settlement in order to reduce unemployment rates and shorten commuting distances, as new details of potentially available local workforce are available. The increasing number of commuters should be taken into account in official statistics in the future. Although the application in this paper refers to NRW, the model is also applicable to countries that perform the LFS and have implemented an FUA structure. Thus, this analysis is transferable to at least all European countries. In Germany, we are facing some limitations in mobile network data. We do not have access to individual signalling data or CDRs. No individual activity movements or changes in individual social behaviour can be used for the estimation. For instance, Toole et al. (2015) have shown that unemployed persons have different mobile phone usage profiles than employed ones. This information may increase the explanatory power in estimating unemployment rates compared to the used distribution of mobile activities over time.

From a methodological point of view, we leave the uncertainty of the difference between the two unemployment rates as further research. So far, we propose an MSE and confidence intervals for each unemployment rate separately. To obtain these two measures for the difference, it is necessary to calculate the covariance between both unemployment rates. For the special case of the difference between a design-based estimator and a FH estimator from the same repeated survey at different points in time, van den Brakel et al. (2016) derives the covariance. It is assumed that the design-based estimator is unbiased and that the covariates for the FH estimator come from the same survey as the design-based estimator. Since these assumptions are not applicable to our case, further research is needed to apply these results to the present case.

In addition, the following research opportunities remain open from an applied perspective. Steele et al. (2017) uses a combination of satellite and mobile phone data to gain more explanatory power in the estimation of poverty indicators. Satellite data include valuable information on a small regional level of building intensities and heights of buildings to differentiate between socially impoverished people, who live in socially weak urban districts, and wealthy people, who are living more likely in less densely populated areas, which could also be suitable for our question. Furthermore, it is of interest to which extend the same differences in unemployment rates also apply to other countries or whether it is a national phenomenon.

Acknowledgments

Würz gratefully acknowledges support by a scholarship of Studienstiftung des deutschen Volkes. The authors are grateful for the computation time provided by the HPC service of the Freie Universität Berlin.

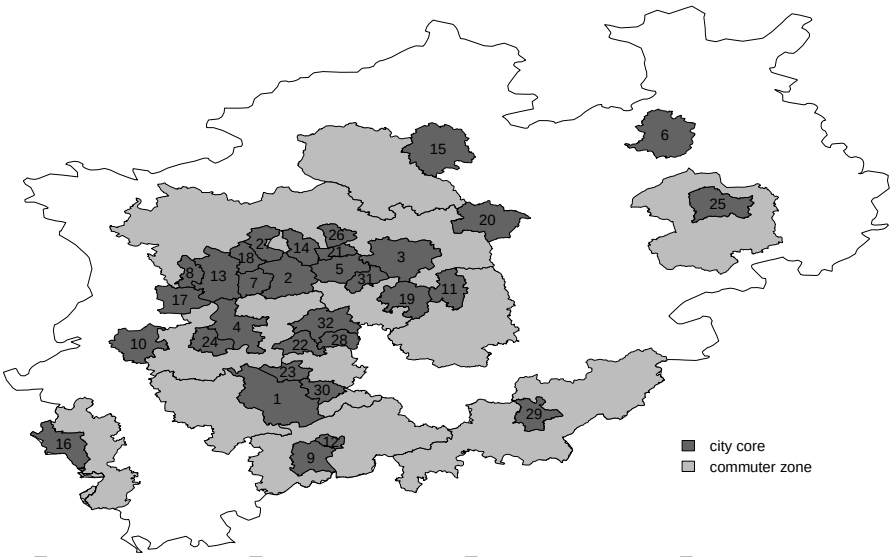
Appendix C

C.1 Mobile network covariates

Table C.1: Mobile network covariates: The last four columns refer to the four different models on unemployment rates at the FUA sublevel. The covariates are based on mobile network data of Deutsche Telekom for the years 2017 and 2018 and represent a statistical week. For each selected variable, the regression coefficient is shown.

Definition of variables		UR_1 male	UR_2 male	UR_1 fem.	UR_2 fem.
Intercept		118.5287	14.8136	0.5674	−14.3924
Proportion of mobile activities of specific subgroup at defined time					
Central European	7 am to 4 pm	-2.6305	-2.7364	-2.2433	
Central European	5 pm to 11 pm	3.1693	4.0103		
Proportion of mobile activities of specific subgroup at defined time on Sunday					
under 50s	8 pm to 11 pm		-0.1478	-0.6384	
20 to 30 year olds	8 pm to 11 pm			0.8168	
Change of mobile activities by nationality from night-time (5 pm to 11 pm) to day-time (7 am to 4 pm)					
African		0.0012	0.0013	−0.0034	−0.0024
Australia Oceania		−0.0001	−0.0001		
Eastern Europe		−0.0680	−0.0836		
North American		−0.0245	−0.0695	−0.0273	−0.0446
Northern Europe		−0.0176	−0.0449		
Southeast Europe		−0.1070	−0.1654	−0.1145	−0.1165
Southern Europe			0.1158	0.1021	
Asia					0.0135
Central Europe					−5.2348
Relative change of mobile activities between two specific times: (time point 1 − time point 2) / time point 2					
10 am	9 pm	−3.2224	4.7858		2.4082
8 pm to 10 pm	9 am to 11 am	3.2963	−3.7265		
4 pm	10 am		−1.2954	−1.1901	
9 am to 11 am	3 am to 5 am			2.4185	
Ratio of mobile activities between two specific times: time point 1 / time point 2					
7 am to 4 pm	5 pm to 11 pm	3.0494		5.2589	
5 pm to 5 am	whole day	−119.1781	−28.0838		
9 am to 11 am	8 pm to 10 pm	−3.9171	−3.5458	−5.5749	
6 am to 4 pm	whole day	−111.7206			30.2304
12 pm to 6 am	7 am to 4 pm		2.7691		
3 am to 5 am	9 am to 11 am			2.0348	

C.2 Map of FUA city cores and commuter zones in NRW



1 Cologne
2 Essen
3 Dortmund
4 Düsseldorf
5 Bochum
6 Bielefeld
7 Mülheim a.d.Ruhr
8 Moers
9 Bonn
10 Mönchengladbach
11 Iserlohn
12 Sankt Augustin
13 Duisburg
14 Gelsenkirchen
15 Münster
16 Aachen
17 Krefeld
18 Oberhausen
19 Hagen
20 Hamm
21 Herne
22 Solingen
23 Leverkusen
24 Neuss
25 Paderborn
26 Recklinghausen
27 Bottrop
28 Remscheid
29 Siegen
30 Bergisch Gladbach
31 Witten
32 Wuppertal

Figure C.1: Assignment of city names to FUA city cores and geographical location of the commuter zones for NRW.

Bibliography

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 (2010). An object-oriented framework for statistical simulation: the R package **simFrame**. Journal of Statistical Software 37(3), 1–36.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. The Annals of Statistics 47(2), 1148–1178.

Augustijn, L. (2018). Berufsbedingte Pendelmobilität, Geschlecht und Stress. `https://www.uni-due.de/imperia/md/content/soziologie/dbsf-2018-02.pdf`. [accessed: 01.2021].

Baran, D. and J. O'Donoghue (2002). Price levels in 2000 for London and the regions compared with the national average. Economic Trends 578, 28–38.

Barton, K. (2018). **MuMIn**: Multi-Model Inference. R package version 1.40.4.

Bates, D., M. Mächler, B. M. Bolker, and S. C. Walker (2015). Fitting linear mixed-effects models using **lme4**. Journal of Statistical Software 67(1), 1–48.

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

Bauer, J. M. and A. Sousa-Poza (2015). Impacts of informal caregiving on caregiver employment, health, and family. Journal of Population Ageing 8(3), 113–145.

Bauer-Hailer, U. (2019). Berufspendler im Bundesländervergleich. `https://www.statistik-bw.de/Service/Veroeff/Monatshefte/PDF/Beitrag19_02_02.pdf`. [accessed: 12.2019].

BBSR (2017). Wachsen und Schrumpfen von Städten und Gemeinden. `https://gis.uba.de/maps/resources/apps/bbsr/index.html?lang=de`. [accessed: 06.2019].

Becker, G. S. (1965). A theory of the allocation of time. The Economic Journal 75(299), 493–517.

Berg, E. and H. Chandra (2014). Small area prediction for a unit-level lognormal model. Computational Statistics & Data Analysis 78, 159–175.

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

Bivand, R. S., E. Pebesma, and V. Gomez-Rubio (2013). Applied Spatial Data Analysis with R. New York: Springer.

BLS (2018). BLS handbook of methods: Chapter 17. The consumer price index. https://www.bls.gov/opub/hom/pdf/homch17.pdf. [accessed: 04.2021].

Boonstra, H. J. (2021). **mcmcsae**: Markov Chain Monte Carlo Small Area Estimation. R package version 0.7.0.

Boonstra, H. J. (2022). **hbsae**: Hierarchical Bayesian Small Area Estimation. R package version 1.2.

Borooah, V. K., P. P. L. McGregor, P. M. McKee, and G. E. Mulholland (1996). Cost of living differences between the regions of the United Kingdom. In J. Hills (Ed.), New Inequalities. The Changing Distribution of Income and Wealth in the United Kingdom, pp. 103 – 132. Cambrige: Cambridge University Press.

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

Breidenbach, J. (2018). **JoSAE**: Unit-Level and Area-Level Small Area Estimation. R package version 0.3.3.

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. In Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: A Methodological Perspective. Statistics Canada.

Brown, M., R. De Haas, and V. Sokolov (2018). Regional inflation, banking integration, and dollarization. Review of Finance 22(6), 2073–2108.

Buch, T., S. Hamann, A. Niebuhr, and A. Rossen (2014). What makes cities attractive? The determinants of urban labour migration in Germany. Urban Studies 51(9), 1960–1978.

Buchanan, J. M. (1991). Opportunity cost. In J. Eatwell, M. Milgate, and P. Newman (Eds.), The World of Economics, pp. 520–525. London: Palgrave Macmillan UK.

Bundesagentur für Arbeit (2022a). Arbeitslosenquote und Unterbeschäftigungsquote. https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Definitionen/Berechnung-der-Arbeitslosenquote/Berechnung-der-Arbeitslosenquote-Nav.html. [accessed: 10.2022].

Bundesagentur für Arbeit (2022b). Pendlerverflechtungen der sozialversicherungspflichtig Beschäftigten nach Ländern - Stichtag: 30.06.2016. https://statistik.arbeitsagentur.de/SiteGlobals/Forms/Suche/Einzelheftsuche_Formular.html?nn=20934&topic_f=beschaeftigung-pendler-blxbl&dateOfRevision=201606-202106. [accessed: 10.2022].

Burgard, J. P. and P. Dörr (2022). Generalized linear mixed models with crossed effects and unit-specific survey weights. Journal of Computational and Graphical Statistics, forthcoming.

Burgard, J. P., R. Münnich, and T. Zimmermann (2016). Impact of sampling designs in small area estimation with applications to poverty measurement. In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation, pp. 85–108. Hoboken: John Wiley & Sons.

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. 379–403. Hoboken: John Wiley & Sons.

Chambers, J. M. and T. J. Hastie (1992). Statistical Models in S. London: Chapman & Hall.

Chambers, R. and H. Chandra (2013). A random effect block bootstrap for clustered data. Journal of Computational and Graphical Statistics 22(2), 452–470.

Chambers, R. and R. Dunstan (1986). Estimating distribution functions from survey data. Biometrika 73(3), 597–604.

Chandra, H. and R. Chambers (2011). Small area estimation under transformation to linearity. Survey Methodology 37(1), 39–51.

Chandra, H. and U. C. Sud (2012). Small area estimation for zero-inflated data. Communications in Statistics-Simulation and Computation 41(5), 632–643.

Chari, A. V., J. Engberg, K. N. Ray, and A. Mehrotra (2015). The opportunity costs of informal elder-care in the United States: new estimates from the American time use survey. Health Services Research 50(3), 871–882.

Charles, K. K. and P. Sevak (2005). Can family caregiving substitute for nursing home care? Journal of Health Economics 24(6), 1174–1190.

Chen, J. and J. Qin (1993). Empirical likelihood estimation for finite populations and the active usage of auxiliary information. Biometrika 80(1), 107–116.

Chen, J., A. M. Variyath, and B. Abraham (2008). Adjusted empirical likelihood and its properties. Journal of Computational and Graphical Statistics 17(2), 426–443.

Choudhry, G. H. and J. N. K. Rao (1989). Small area estimation using models that combine time series and cross sectional data. In A. C. Singh and P. Whitridge (Eds.), Proceedings of Statistics Canada Symposium on Analysis of Data in Time, pp. 67–74. Ottawa: Statistics Canada.

Costa, A., A. Satorra, and E. Ventura (2006). Improving small area estimation by combining surveys: new perspectives in regional statistics. Statistics and Operations Research Transactions 30(1), 101–121.

Dagdoug, M., C. Goga, and D. Haziza (2022). Model-assisted estimation through random forests in finite population sampling. Journal of the American Statistical Association, forthcoming.

Datta, G. S. 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.

De Meersman, F., G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis, and H. I. Reuter (2016). Assessing the quality of mobile phone data as a source of statistics. https://ec.europa.eu/eurostat/cros/system/files/assessing_the_quality_of_mobile_phone_data_as_a_source_of_statistics_q2016.pdf. [accessed: 11.2018].

Defra and ONS (2019). Living costs and food survey, 2008-2014. https://doi.org/10.5255/UKDA-SN-7992-4 (DOI for 2014 only). 3rd Edition. UK Data Service. Data collection. SN: 7992 and also SN: 6385, 6655, 6945, 7272, 7472, 7702.

Diallo, M. S. and J. N. K. Rao (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. Scandinavian Journal of Statistics 45(4), 1092–1116.

Dijkstra, L. and H. Poelman (2011). Archive:European cities - the EU-OECD functional urban area definition. https://ec.europa.eu/eurostat/statistics-explained/index.php/Archive:European_cities_%E2%80%93_the_EU-OECD_functional_urban_area_definition#A_harmonised_definition. [accessed: 06.2019].

Duan, N. (1983). Smearing estimate: a nonparametric retransformation method. Journal of the American Statistical Association 78(383), 605–610.

Duong, T. (2022). **ks**: Kernel Smoothing. R package version 1.13.4.

Duran, H. E. (2016). Inflation differentials across regions in Turkey. The South East European Journal of Economics and Business 11(1), 7–17.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. Econometrica 71(1), 355–364.

Elhorst, P. J. (2003). The mystery of regional unemployment differentials: Theoretical and empirical explanationa. Journal of Economic Surveys 17(5), 709–748.

Emerson, S. C. and A. B. Owen (2009). Calibration of the empirical likelihood method for a vector mean. Electronic Journal of Statistics 3, 1161–1192.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications 14(1), 153–158.

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

Esteban, M. D., D. Morales, and A. Pérez (2016). Area-level spatio-temporal small area estimation models. In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation, pp. 205–226. Hoboken: John Wiley & Sons.

Esteban, M. D., D. Morales, A. Pérez, and L. Santamaía (2011). Two area-level time models for estimating small area poverty indicators. Journal of the Indian Society of Agricultural Statistics 66(11), 75–89.

EU regulation (1998). EU regulation 1687/98 on HICP. https:\www.eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31998R1687&from=EN. [accessed: 04.2021].

European Parliament and Council (2003). Regulation (EC) No 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics (NUTS). Official Journal of the European Union 154(1), 1–41.

Eurostat (2017). City statistics (urb): national reference metadata in euro SDMX metadata structure (ESMS). https://ec.europa.eu/eurostat/cache/metadata/EN/urb_esms_de.htm. [accessed: 11.2018].

Eurostat (2018a). Dataset details: harmonised unemployment rate by sex. https://ec.europa.eu/eurostat/web/products-datasets/-/teilm020. [accessed: 10.2018].

Eurostat (2018b). Glossary:Functional urban area. https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Functional_urban_area. [accessed: 05.2018].

Eurostat (2018c). NUTS - Systematik der Gebietseinheiten für die Statistik: Hintergrund. https://ec.europa.eu/eurostat/web/nuts/background. [accessed: 08.2019].

Eurostat (2019a). DataCollection: Precision Level DCF. https://datacollection.jrc.ec.europa.eu/wordef/precision-level-dcf. [accessed: 06.2019].

Eurostat (2019b). EU labour force survey database user guide. https://ec.europa.eu/eurostat/documents/1978984/6037342/EULFS-Database-UserGuide.pdf. [accessed: 08.2019].

Eurostat (2019c). Labour force survey in the EU, candidate and EFTA countries - main characterstics of national surveys, 2018. https://ec.europa.eu/eurostat/de/web/products-statistical-reports/-/KS-FT-19-008?inheritRedirect=true. [accessed: 01.2021].

Eurostat (2019d). Städte (Urban Audit): Datenbank. https://ec.europa.eu/eurostat/de/web/cities/data/database. [accessed: 08.2019].

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.

Fengki, A. O., K. A. Notodiputro, and K. Sadik (2020). Bisakah memperoleh statistik indeks harga konsumen tingkat provinsi di indonesia dengan ketelitian yang lebih baik? In B. F. Wiratama (Ed.), Seminar Nasional Official Statistics, pp. 297–306. Jakarta: Diterbitkan oleh Politeknik Statistika STIS.

Fenwick, D. and J. O'Donoghue (2003). Developing estimates of relative regional consumer price levels. Economic Trends 599, 72–83.

Franconi, L., D. Ichim, M. D'Alò, and S. Cruciani (2017). Guidelines for labour market area delineation process: from definition to dissemination. https://ec.europa.eu/eurostat/cros/system/files/guidelines_for_lmas_production08082017_rev300817.pdf. [accessed: 01.2021].

Frick, J. R. and J. Goebel (2008). Regional income stratification in unified Germany using a Gini decomposition approach. Regional Studies 42(4), 555–577.

Fuchs-Schündeln, N., D. Krueger, and M. Sommer (2010). Inequality trends for Germany in the last two decades: a tale of two countries. Review of Economic Dynamics 13(1), 103–132.

Gajewski, P. (2017). Sources of regional inflation in Poland. Eastern European Economics 55(3), 261–276.

Galiana, L., B. Sakarovitch, and Z. Smoreda (2018). Understanding socio-spatial segregation in French cities with mobile phone data. http://www.dgins2018.ro/wp-content/uploads/2018/10/08-FR-dgins_segregation_1800905.pdf. [accessed: 11.2018].

Gans, P. (2017). Urban population development in Germany (2000-2014): the contribution of migration by age and citizenship to reurbanisation. Comparative Population Studies 42, 319–352.

Goebel, J., M. M. Grabka, S. Liebig, M. Kroh, D. Richter, C. Schröder, and J. Schupp (2019). The German Socio-Economic Panel (SOEP). Jahrbücher für Nationalökonomie und Statistik 239(2), 345–360.

González-Manteiga, W., M. J. Lombardía, I. Molina, D. Morales, and L. Santamaría (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. Computational Statistics & Data Analysis 52(12), 5242–5252.

Görzig, B., M. Gornig, and A. Werwatz (2008). Firm wage differentiation in Eastern Germany: a non-parametric analysis of the wage spread. Economics of Transition 16(2), 273–292.

Graf, M., J. M. Marín, and I. Molina (2019). A generalized mixed model for skewed distributions applied to small area estimation. TEST 28(2), 565–597.

Grözinger, G. (2009). Achtung Lebensgefahr! Indirekte Effekte regionaler Arbeitslosigkeit auf Lebensweise und -qualität. European Journal of Economics and Economic Policies: Intervention 6(1), 12–24.

Grözinger, G. (2018). Regionale Arbeitslosigkeit: Falsche Eindrücke von Stadt-Land-Differenzen. Wirtschaftsdienst 98(1), 68–70.

Guadarrama, M., I. Molina, and J. N. K. Rao (2018). Small area estimation of general parameters under complex sampling designs. Computational Statistics & Data Analysis 121, 20 – 40.

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. (2018). Use of mobile phone data for official statistics. METHODS - APPROACHES - DEVELOPMENTS: Information of the German Federal Statistical Office 1(2), 6–9.

Hadam, S. (2021). Pendler Mobil: Die Verwendung von Mobilfunkdaten zur Unterstützung der amtlichen Pendlerstatistik. AStA Wirtschafts- und Sozialstatistisches Archiv 15(3), 197–235.

Hadam, S., N. Würz, and A.-K. Kreutzmann (2020). Estimating regional unemployment with mobile network data for functional urban areas in Germany. Preprint: `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 Mirco-data. Luxembourg: Office for the Official Publications of the European Communities.

Hajjem, A., F. Bellavance, and D. Larocque (2014). Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation 84(6), 1313–1328.

Hall, P. and T. Maiti (2006). On parametric bootstrap methods for small area prediction. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(2), 221–238.

Han, P. and J. F. Lawless (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. Statistica Sinica 29(3), 1321–1342.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

Hayes, P. (2005). Estimating UK regional price indices, 1974–96. Regional Studies 39(3), 333–344.

ILO (2018). Unemployment rate. http://www.ilo.org/ilostat-files/Documents/description_UR_EN.pdf. [accessed: 10.2018].

IT.NRW (2019). Berufspendler 2011 – 2018 nach Pendlerart, Beschäftigungsumfang und Geschlecht. https://www.it.nrw/statistik/eckdaten/berufspendler-2011-2018-nach-pendlerart-beschaeftigungsumfang-und-geschlecht. [accessed: 12.2019].

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

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

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

Karlberg, F. (2000). Population total prediction under a lognormal superpopulation model. METRON 58(3/4), 53–80.

Klammer, U. and K. Menke (2020). Gender-Datenreport. https://www.bpb.de/izpb/307413/geschlechterdemokratie. [accessed: 01.2021].

Kohn, K. and D. Antonczyk (2013). The aftermath of reunification: sectoral transition, gender and rising wage inequality in East Germany. Economics of Transition 21(1), 73–110.

Kosfeld, R. and C. Dreger (2006). Thresholds for employment and unemployment. A spatial analysis of German regional labour markets 1999-2000. Papers in Regional Science 85(4), 523–542.

Kosfeld, R., H.-F. Eckey, and J. Lauridsen (2008). Disparities in prices and income across German NUTS 3 regions. Applied Economics Quarterly 54(2), 123–141.

Kosfeld, R., H.-F. Eckey, and M. Schüßler (2009). Ökonometrische Messung regionaler Preisniveaus auf der Basis örtlich beschränkter Erhebungen. German Council for Social and Economic Data (RatSWD) Research Notes 33, 1–33.

KOSIS-Gemeinschaft Urban Audit (2013). Das deutsche Urban Audit - Städtevergleich im Europäischen Statistischen System. http://www.staedtestatistik.de/fileadmin/urban-audit/UA_Broschuere_2013.pdf. [accessed: 08.2019].

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(5), 1865–1894.

Krennmair, P., N. Würz, and T. Schmid (2022). Analysing opportunity cost of care work using mixed effects random forests under aggregated census data. Preprint: https://arxiv.org/abs/2204.10736.

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

Kroh, M., S. Kühne, R. Siegers, and V. Belcheva (2018). SOEP-core-documentation of sample sizes and panel attrition (1984 until 2016). SOEP Survey Papers - Series C - Data Documentations 480, 1–91.

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

Li, H., Y. Liu, and R. Zhang (2019). Small area estimation under transformed nested-error regression models. Statistical Papers 60(4), 1397–1418.

Lumley, T. (2004). Analysis of complex survey samples. Journal of Statistical Software 9(1), 1–19.

Maechler, M., W. Stahel, A. Ruckstuhl, C. Keller, A. Hauser, C. Buser, L. Gygax, B. Venables, T. Plate, I. Flückiger, M. Wolbers, M. Keller, S. Dudoit, J. Fridlyand, G. Snow, H. A. Nielsen, V. Carey, B. Bolker, P. Grosjean, F. Ibanez, C. Savi, C. Geyer, and J. Oehlschlägel (2021). **sfsmisc**: Utilities from 'Seminar fuer Statistik' ETH Zurich. R package version 1.1-12.

Marchetti, S., G. Bertarelli, L. Biggeri, G. Giusti, M. Pratesi, and F. Schirripa-Spagnolo (2019). Small area poverty indicators adjusted using local price indexes. https://www.centrodagum.it/wp-content/uploads/2019/07/Presentation_Pratesi.pdf. [accessed: 04.2021].

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.

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

Marhuenda, Y., I. Molina, and D. Morales (2013). Small area estimation with spatio-temporal Fay-Herriot models. Computational Statistics & Data Analysis 58, 308–325.

Marino, M. F., M. G. Ranalli, N. Salvati, and M. Alfò (2019). Semiparametric empirical best prediction for small area estimation of unemployment indicators. Annals of Applied Statistics 13(2), 1166–1197.

Martini, A. and S. Loriga (2017). Small area estimation of employment and unemployment for local labour market areas in Italy. https://www.dst.dk/ext/4299453410/0/formid/4-2-Small-Area-Estimation-of-employment-and-unemployment-for-Local_Labour-Market-Areas-in-Italy--pdf. [accessed: 11.2018].

Mendez, G. and S. Lohr (2011). Estimating residual variance in random forest regression. Computational Statistics & Data Analysis 55(11), 2937–2950.

Möbert, J. (2018). The German housing market in 2018. https://everydaypoint.com/wp-content/uploads/2019/04/The_German_housing_market_in_2018-1.pdf. [accessed: 09.2019].

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

Molina, I. and N. Martín (2018). Empirical best prediction under a nested error model with log transformation. The Annals of Statistics 46(5), 1961–1993.

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

Molina, I. and E. Strzalkowska-Kominiak (2020). Estimation of proportions in small areas: application to the labour force using the Swiss census structural survey. Journal of the Royal Statistical Society: Series A (Statistics in Society) 183(1), 281–310.

Mudrazija, S. (2019). Work-related opportunity costs of providing unpaid family care in 2013 and 2050. Health Affairs 38(6), 1003–1010.

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

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

Ochalek, J., J. Lomas, and K. Claxton (2018). Estimating health opportunity costs in low-income and middle-income countries: a novel approach and evidence from cross-country data. BMJ Global Health 3(6), 1–10.

O'Donoghue, J. (2017). The effect of variance in the weights on the CPI and RPI. Survey Methodology Bulletin 77, 1–27.

OECD (2020). Working age population (indicator). OECD, Paris. https://www.oecd-ilibrary.org/social-issues-migration-health/working-age-population/indicator/english_d339918b-en. [accessed: 04.2021].

Oliva-Moreno, J., L. M. Peña Longobardo, L. García-Mochón, M. del Río Lozano, I. Mosquera Metcalfe, and M. d. M. García-Calvente (2019). The economic value of time of informal care and its determinants (the CUIDARSE study). PLOS ONE 14(5), 1–15.

ONS (2011). UK relative regional consumer price levels for goods and services for 2010. https://www.ons.gov.uk/ons/rel/cpi/regional-consumer-price-levels/2010/uk-relative-regional-consumer-price-levels-for-goods-and-services-for-2010.pdf. [accessed: 04.2021].

ONS (2014). Consumer prices indices technical manual. https://www.ons.gov.uk/ons/rel/cpi/consumer-price-indices---technical-manual/2014/index.html. [accessed: 04.2021].

ONS (2018). Development of regional household expenditure measures. https://www.ons.gov.uk/economy/regionalaccounts/grossdisposablehouseholdincome/articles/developmentofregionalhouseholdexpendituremeasures/latest. [accessed: 04.2021].

ONS (2019). Consumer prices indices technical manual. https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/consumerpriceindicescpiandretailpricesindexrpiitemindicesandpricequotes. [accessed: 04.2021].

ONS (2020). Consumer price inflation item indices and price quotes. https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/consumerpriceindicescpiandretailpricesindexrpiitemindicesandpricequotes. [accessed: 04.2021].

Opsomer, J. D., G. Claeskens, M. G. Ranalli, G. Kauermann, and F. J. Breidt (2008). Nonparametric small area estimation using penalized spline regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(1), 265–286.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. The Annals of Statistics 18(1), 90–120.

Owen, A. B. (2001). Empirical Likelihood. New York: Chapman and Hall/CRC.

Parzen, E. (1962). On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3), 1065–1076.

Patuelli, R., D. A. Griffitha, M. Tiefelsdorf, and P. Nijkamp (2011). Spatial filtering and eigenvector stability: Space-time models for German unemployment data. International Regional Science Review 34(2), 253–280.

Pereira, L. N., J. Mendes, and P. S. Coelho (2011). Estimation of unemployment rates in small areas of Portugal: a best linear unbiased prediction approach versus a hierarchical Bayes approach. In 17th European Young Statisticians Meeting. Universidade Nova de Lisboa.

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

Pfeffermann, D., C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash (1998). Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60(1), 23–40.

Pfeffermann, D. and M. Sverchkov (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. Journal of the American Statistical Association 102(480), 1427–1439.

Pfeffermann, D., B. Terryn, and F. A. S. Moura (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. Survey Methodology 34(2), 235–249.

Pinheiro, J., D. Bates, and R Core Team (2022). **nlme**: Linear and Nonlinear Mixed Effects Models. R package version 3.1-155.

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

Purwono, R., M. Z. Yasin, and M. K. Mubin (2020). Explaining regional inflation programmes in Indonesia: does inflation rate converge? Economic Change and Restructuring 53(4), 571–590.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. The Annals of Statistics 22(1), 300 – 325.

Rabe-Hesketh, S. and A. Skrondal (2006). Multilevel modelling of complex survey data. Journal of the Royal Statistical Society: Series A (Statistics in Society) 169(4), 805–827.

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

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

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

Rienzo, C. (2017). Real wages, wage inequality and the regional cost-of-living in the UK. Empirical Economics 52(4), 1309–1335.

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.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics 27(3), 832–837.

RPI Advisory Committee (1971). Proposals for retail prices indices for regions. London: H.M. Stationery Office.

Säfken, B., D. Rügamer, T. Kneib, and S. Greven (2021). Conditional model selection in mixed-effects models with **cAIC4**. Journal of Statistical Software 99(8), 1–30.

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

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

Schoch, T. (2014). **rsae**: Robust Small Area Estimation. R package version 0.1-5.

Scott, D. W. (2015). Multivariate Density Estimation: Theory, Practice, and Visualization. Hoboken: John Wiley & Sons.

Sexton, J. and P. Laake (2009). Standard errors for bagged and random forest estimators. Computational Statistics & Data Analysis 53(3), 801–811.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 53(3), 683–690.

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

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

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

Stanfors, M., J. C. Jacobs, and J. Neilson (2019). Caregiving time costs and trade-offs: gender differences in Sweden, the UK, and Canada. SSM - Population Health 9, 100501.

Statistics Bureau of Japan (2020). Regional CPIs for Japan. https://www.e-stat.go.jp/en/stat-search/files?page=1&layout=datalist&toukei=00200573&tstat=000001084976&cycle=1&year=20200&month=11010303&tclass1=000001085955&tclass2val=0. [accessed: 04.2021].

Statistisches Bundesamt (2015). Zensus 2011 Methoden und Verfahren. Statistisches Bundesamt, Wiesbaden. https://www.zensus2011.de/SharedDocs/Downloads/D

E/Publikationen/Aufsaetze_Archiv/2015_06_MethodenUndVerfahren.pdf?__blob=publicationFile&v=6. [accessed: 12.2020].

Statistisches Bundesamt (2018). Preise Verbraucherpreisindex für Deutschland Qualitätsbericht. Statistisches Bundesamt, Wiesbaden. www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Preise/verbraucherpreis.pdf?__blob=publicationFile. [accessed: 04.2021].

Statistisches Bundesamt (2020). Regional CPIs for Germany. Statistisches Bundesamt, Wiesbaden. https://www-genesis.destatis.de/genesis/online (search '61111-0011'). [accessed: 04.2021].

Statistisches Bundesamt (2021). Registered unemployed, unemployment rate by sex. Statistisches Bundesamt, Wiesbaden. https://www.destatis.de/EN/Themes/Labour/Labour-Market/Unemployment/Tables/lrarb002.html. [accessed: 01.2021].

Statistisches Bundesamt (2022). Equipment of households with information and communication technology (Germany). https://www.destatis.de/EN/Themes/Society-Environment/Income-Consumption-Living-Conditions/Equipment-Consumer-Durables/Tables/liste-equipment-households-information-communication-technology-germany.html#55714. [accessed: 10.2022].

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

Sugasawa, S. (2016). **rhnerm**: Random Heteroscedastic Nested Error Regression. R package version 1.1.

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

Sugasawa, S. and T. Kubokawa (2019). Adaptively transformed mixed-model prediction of general finite-population parameters. Scandinavian Journal of Statistics 46(4), 1025–1046.

Swanson, D. C., S. K. Hauge, M. L. Schmidt, et al. (1999). Evaluation of composite estimation methods for cost weights in the CPI. In Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington D.C. American Statistical Association.

Tillmann, P. (2013). Inflation targeting and regional inflation persistence: evidence from Korea. Pacific Economic Review 18(2), 147–161.

Toole, J. L., Y.-R. Lin, E. Muehlegger, D. Shoag, M. C. González, and D. Lazer (2015). Tracking employment shocks using mobile phone data. Journal of the Royal Society Interface 12(107), 20150185.

Truskinovsky, Y. and N. Maestas (2018). Caregiving and labor force participation: new evidence from the American time use survey. Innovation in Aging 2(1), 580.

Tzavidis, N., S. Marchetti, and R. Chambers (2010). Robust estimation of small-area means and quantiles. Australian & New Zealand Journal of Statistics 52(2), 167–186.

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.

UK Statistics Authority (2013). Statistics on consumer price inflation - assessment report 257. https://www.osr.statisticsauthority.gov.uk/wp-content/uploads/2015/11/images-assessmentreport257statisticsonconsumerpriceinflatio_tcm97-43135.pdf. [accessed: 04.2021].

U.S. Bureau of Labor Statistics (2021). Labor force statistics from the current population survey: Concepts and definitions. https://www.bls.gov/cps/definitions.htm#:~:text=The%20unemployment%20rate%20represents%20the,%C3%B7%20Labor%20Force)%20x%20100. [accessed: 10.2022].

van den Brakel, J. A., B. Buelens, and H.-J. Boonstra (2016). Small area estimation to quantify discontinuities in repeated sample surveys. Journal of the Royal Statistical Society: Series A (Statistics in Society) 179(1), 229–250.

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

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

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113(523), 1228–1242.

Wager, S., T. Hastie, and B. Efron (2014). Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. The Journal of Machine Learning Research 15(1), 1625–1651.

Walter, P., M. Groß, T. Schmid, and N. Tzavidis (2021). Domain prediction with grouped income data. Journal of the Royal Statistical Society: Series A (Statistics in Society) 184(4), 1501–1523.

Weber, A. A. and G. W. Beck (2005). Price stability, inflation convergence and diversity in EMU: does one size fit all? CFS Working Paper 2005(30), Goethe University Frankfurt, Center for Financial Studies, Frankfurt.

Wickham, H. (2016). **ggplot2**: Elegant Graphics for Data Analysis. New York: Springer.

Wingfield, D., D. Fenwick, and K. Smith (2005). Relative regional consumer price levels in 2004. Economic Trends 615, 36–45.

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), 1–17.

Würz, N. (2022). **saeTrafo**: Transformations for Unit-Level Small Area Models. R package version 1.0.0.

Würz, N., T. Schmid, and N. Tzavidis (2022). Estimating regional income indicators under transformations and access to limited population auxiliary information. Journal of the Royal Statistical Society: Series A (Statistics in Society), forthcoming.

Yang, L. (1995). Transformation-density estimation. Ph. d. thesis, University of North Carolina, Chapel Hill.

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

Yesilyurt, F. and J. P. Elhorst (2014). A regional analysis of inflation dynamics in Turkey. The Annals of regional science 52(1), 1–17.

You, Y. and J. N. K. Rao (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. Canadian Journal of Statistics 30(3), 431–439.

Zhang, H., J. Zimmerman, D. Nettleton, and D. J. Nordman (2019). Random forest prediction intervals. The American Statistician 74(4), 392–406.

Zimmermann, T., F. Polidoro, F. D. Leo, M. Fedeli, J. Burger, J. van den Brakel, M. Pratesi, C. Giusti, S. Marchetti, L. Biggeri, G. Bertarelli, F. S. Spagnolo, T. Laureti, I. Benedetti, P. A. Smith, J. Dawber, N. Tzavidis, A. Luna, J. O'Donoghue, T. Flower, H. Thomas, N. Würz, T. Schmid, C. Articus, J. P. Burgard, C. Caratiola, H. Dieckmann, F. Ertz, J. Krause, R. Münnich, and A.-L. Wölwer. (2020). Regional poverty measurement as a prototype for modern indicator methodology: guidelines for best practices implementation for transferring methodology. MAKSWELL. `https://www.makswell.eu/attached_documents/output_deliverables/deliverable_3.2.pdf`. [accessed: 04.2021].

Summaries

Abstracts in English

Abstract: Estimating regional income indicators under transformations and access to limited population auxiliary information

Spatially disaggregated income indicators are typically estimated by using model-based methods that assume access to auxiliary information from population micro-data. In many countries like Germany and the UK population micro-data are not publicly available. In this work we propose small area methodology when only aggregate population-level auxiliary information is available. We use data-driven transformations of the response to satisfy the parametric assumptions of the used models. In the absence of population micro-data, appropriate bias-corrections for small area prediction are needed. Under the approach we propose in this paper, aggregate statistics (means and covariances) and kernel density estimation are used to resolve the issue of not having access to population micro-data. We further explore the estimation of the mean squared error using the parametric bootstrap. Extensive model-based and design-based simulations are used to compare the proposed method to alternative methods. Finally, the proposed methodology is applied to the 2011 Socio-Economic Panel and aggregate census information from the same year to estimate the average income for 96 regional planning regions in Germany.

Keywords: Census, density estimation, official statistics, unit-level models, small area estimation

Abstract: The R package saeTrafo for estimating unit-level small area models under transformations

The R package **saeTrafo** provides new statistical methodology for the estimation of small area means using unit-level models under transformations. The method of Würz et al. (2022) enables the use of unit-level models dealing with both limited auxiliary data (often the only source of data due to confidentiality agreements) and skewed distributed dependent variables like income (by using transformations such as the log or data-driven log-shift). In addition to the implementation of the new methodology, **saeTrafo** provides established methods for unit-level models under transformations, allowing further applications and comparisons. It is of advantage that the most suitable method is automatically selected and uncertainty estimates are easily offered. In addition, tools for creating plots (model validation and estimator evaluation), visualisation on maps and exporting to Excel and OpenDocument `Spreadsheets`

are provided. The functionalities of the package are demonstrated with exemplary data based on Austrian income and living conditions.
Keywords: Official statistics, survey statistics, small area estimation, nested error regression model, transformations

Abstract: Analysing opportunity cost of care work using mixed effects random forests under aggregated census data

Reliable estimators of the spatial distribution of socio-economic indicators are essential for evidence-based policy-making. As sample sizes are small for highly disaggregated domains, the accuracy of the direct estimates is reduced. To overcome this problem small area estimation approaches are promising. In this work we propose a small area methodology using machine learning methods. The semi-parametric framework of mixed effects random forest combines the advantages of random forests (robustness against outliers and implicit model-selection) with the ability to model hierarchical dependencies. Existing random forest-based methods require access to auxiliary information on population-level. We present a methodology that deals with the lack of population micro-data. Our strategy adaptively incorporates aggregated auxiliary information through calibration-weights - based on empirical likelihood - for the estimation of area-level means. In addition to our point estimator, we provide a non-parametric bootstrap estimator measuring its uncertainty. The performance of the proposed point estimator and its uncertainty measure is studied in model-based simulations. Finally, the proposed methodology is applied to the 2011 Socio-Economic Panel and aggregate census information from the same year to estimate the average opportunity cost of care work for 96 regional planning regions in Germany.
Keywords: Official statistics, small area estimation, mean squared error, tree-based methods

Abstract: Experimental UK regional consumer price inflation with model-based expenditure weights

Like many other countries, the United Kingdom (UK) produces a national consumer price index (CPI) to measure inflation. Presently, CPI measures are not produced for regions within the UK. It is believed that, using only available data sources, a regional CPI would not be precise or reliable enough as an official statistic, primarily because the regional partitioning of the data makes sample sizes too small. We investigate this claim by producing experimental regional CPIs using publicly available price data, and deriving expenditure weights from the Living Costs and Food survey. We detail the methods and challenges of developing a regional CPI and evaluate its reliability. We then assess whether model-based methods such as smoothing and small area estimation significantly improve the measures. We find that a regional CPI can be produced with available data sources, however it appears to be excessively volatile over time, mainly due to the weights. Smoothing and small area estimation improve the reliability of the regional CPI series to some extent but they remain too volatile for regional policy use. This research provides a valuable framework for the development of a more viable regional CPI measure for the UK in the future.

Keywords: CPI conceptual framework, basket of goods and services, small area estimation, Fay-Herriot models

Abstract: Estimating regional unemployment with mobile network data for functional urban areas in Germany

The ongoing growth of cities due to better job opportunities is leading to increased labour-related commuter flows in several countries. On the one hand, an increasing number of people commute and move to the cities, but on the other hand, the labour market indicates higher unemployment rates in urban areas than in the surrounding areas. We investigate this phenomenon on regional level by an alternative definition of unemployment rates in which commuting behaviour is integrated. We combine data from the Labour Force Survey with dynamic mobile network data by small area models for the federal state North Rhine-Westphalia in Germany. From a methodical perspective, we use a transformed Fay-Herriot model with bias correction for the estimation of unemployment rates and propose a parametric bootstrap for the mean squared error estimation that includes the bias correction. The performance of the proposed methodology is evaluated in a case study based on official data and in model-based simulations. The results in the application show that unemployment rates (adjusted by commuters) in German cities are lower than traditional official unemployment rates indicate.
Keywords: Bias correction, Fay-Herriot model, mean squared error, small area estimation, unemployment rates

Kurzzusammenfassungen auf Deutsch

Zusammenfassung: Schätzung regionaler Einkommensindikatoren unter Transformationen und limitiertem Zugang zu Hilfsinformationen aus der Population

Kleinräumige Einkommensindikatoren werden meist mit modellbasierten Methoden geschätzt, die Hilfsinformationen über die Population auf Mikrodaten-Ebene benötigen. In zahlreichen Ländern, wie Deutschland und dem Vereinigten Königreich, sind Populations-Mikrodaten jedoch nicht öffentlich zugänglich. In dieser Arbeit werden Small-Area-Methoden vorgeschlagen, die bei ausschließlicher Verfügbarkeit von aggregierten Populations-Hilfsinformationen verwendet werden können. Um die parametrischen Modellannahmen zu erfüllen, wird die abhängige Variable mittels datengetriebener Transformation angepasst. Hierbei werden geeignete Verzerrungs-Korrekturen für die Small-Area-Vorhersagen benötigt. Der vorgeschlagene Ansatz kombiniert aggregierte Statistiken (Mittelwerte und Kovarianzen) und Kerndichte-Schätzungen, um das Problem des fehlenden Zugangs zu Populations-Mikrodaten zu adressieren. Zudem wird die Schätzung des mittleren quadratischen Fehlers mittels parametrischem Bootstrap-Verfahren vorgestellt. Ausführliche modellbasierte und designbasierte Simulationen werden verwendet, um die vorgeschlagene Methode mit alternativen Methoden zu vergleichen. Abschließend wird die Methode auf das Sozioökonomische Panel von 2011 unter Verwendung von aggregierten Zensusdaten aus demselben Jahr angewandt, um das durchschnittliche Einkommen für die 96 deutschen Raumordnungsregionen zu schätzen.

Schlüssewörter: Zensus, Dichteschätzung, Amtliche Statistik, Unit-Level-Modelle, Small-Area-Schätzung

Zusammenfassung: Das R Paket saeTrafo zur Schätzung von Unit-Level Small-Area-Modellen unter Transformationen

Das R-Paket **saeTrafo** stellt eine neue statistische Methode zur Schätzung von Small-Area-Mittelwerten unter Verwendung von Unit-Level-Modellen mit Transformationen zur Verfügung. Die Methode von Würz et al. (2022) ermöglicht die Anwendung von Unit-Level-Modellen unter limitierten Hilfsinformationen (aufgrund von Datenschutzverpflichtungen oft die einzige Datenquelle) für schief verteilte abhängige Variablen wie zum Beispiel Einkommen (mittels Log- oder datengetriebener Log-Shift-Transformation). Zusätzlich zur Implementation der neuen Methode stellt **saeTrafo** etablierte Methoden für Unit-Level-Modelle unter Transformationen bereit, sodass weitere Anwendungen und Vergleiche ermöglicht werden. Dabei profitiert der Nutzer von der automatischen Auswahl der geeigneten Methode und der direkten Bereitstellung von Unsicherheitsschätzern. Zusätzlich werden die Erstellung von Plots (Modellvalidierung und Bewertung der Schätzer), die Visualisierung auf Karten und der Export nach Excel- und OpenDocument-`Spreadsheets` ermöglicht. Die Funktionalitäten des Paketes werden anhand von beispielhaften Daten zu österreichischen Einkommens- und Lebensbedingungen demonstriert.
Schlüsselwörter: Amtliche Statistik, Survey-Statistik, Small-Area-Schätzung, verschachteltes Fehlerregressionsmodell, Transformationen

Zusammenfassung: Analyse der Opportunitätskosten von Pflegearbeit mit Mixed-Effects-Random-Forests unter Verwendung aggregierter Zensusdaten

Für evidenzbasierte politische Entscheidungsfindungen sind zuverlässige Schätzungen der räumlichen Verteilung sozioökonomischer Indikatoren unerlässlich. Da höhere räumliche Auflösungen mit kleineren Stichprobengrößen einhergehen, ist die Genauigkeit der direkten Schätzer reduziert. Um dieses Problem zu lösen, sind Small-Area-Verfahren vielversprechend. Diese Arbeit schlägt eine Small-Area-Methode vor, die Machine-Learning-Verfahren verwendet. Das semiparametrische Konzept von Mixed-Effects-Random-Forests kombiniert die Vorteile von Random-Forests (Robustheit gegenüber Ausreißern und implizite Modellauswahl) mit der Fähigkeit hierarchische Abhängigkeiten zu modellieren. Allerdings benötigen Random-Forest-Methoden Zugang zu Hilfsinformationen auf Populations-Ebene. Daher wird eine Methode vorgestellt, die mit fehlenden Populations-Mikrodaten umgehen kann. Die Strategie beruht auf dem adaptiven Einbezug - basierend auf der empirischen Likelihood - von aggregierten Hilfsinformationen in die Kalibrierungsgewichte für die Schätzung von Mittelwerten auf Gebietsebene. Zusätzlich zu dem Punktschätzer wird ein nicht-parametrischer Bootstrap-Schätzer als Unsicherheitsmaß bereitgestellt. Die Qualität des vorgeschlagenen Punktschätzers sowie dessen Unsicherheitsmaß wird in modellbasierten Simulationen untersucht. Abschließend wird die vorgeschlagene Methode auf das Sozioökonomische Panel von 2011 unter Verwendung von aggregierten Zensusdaten aus demselben Jahr angewandt, um die durchschnittlichen Opportunitätskosten für Pflegearbeit in den 96 deutschen Raumordnungsregionen

zu schätzen.
Schlüsselwörter: Amtliche Statistik, Small-Area-Schätzung, mittlere quadratische Abweichung, baumbasierte Verfahren

Zusammenfassung: Experimentelle regionale Verbraucherpreisinflation für UK mittels modellbasierten Ausgabenanteilen

Wie eine Vielzahl von Ländern bestimmt das Vereinigte Königreich (UK) einen nationalen Verbraucherpreisindex (VPI) zur Messung der Inflation. Allerdings werden gegenwärtig keine VPI-Berechnungen auf Ebene der Regionen bereitgestellt. Es wird angenommen, dass ein regionaler VPI aus den verfügbaren Datenquellen nicht genau bzw. zuverlässig genug für eine amtliche Statistik ist, da insbesondere die Stichprobengrößen durch die regionale Untergliederung zu klein werden. Diese Annahme wird durch die Konstruktion experimenteller regionaler VPIs unter Verwendung öffentlich zugänglicher Preisdaten und der Bestimmung von Ausgabenanteilen aus der Einkommens- und Verbrauchsstichprobe untersucht. Es wird auf die Methoden und Herausforderungen beim Erstellen regionaler VPIs eingegangen sowie ihre Zuverlässigkeit bewertet. Anschließend wird untersucht, ob modellbasierte Methoden (Glättung und Small-Area-Schätzungen) die Messwerte verbessern. Es zeigt sich, dass ein regionaler VPI mit den verfügbaren Datenquellen erstellt werden kann, jedoch scheint er im Zeitverlauf übermäßig volatil zu sein, was hauptsächlich auf die Ausgabenanteile zurückzuführen ist. Glättung und Small-Area-Schätzung verbessern die Zuverlässigkeit der regionalen VPI-Reihen bis zu einem gewissen Grad, aber sie bleiben dennoch zu unbeständig für die Verwendung in der Regionalpolitik. Demnach bietet diese Untersuchung einen wertvollen Startpunkt für die zukünftige Entwicklung eines tragfähigeren regionalen VPIs für UK.
Schlüsselwörter: Konzeptueller Rahmen des VPIs, Warenkorb, Small-Area-Schätzung, Fay-Herriot-Modelle

Zusammenfassung: Schätzung von regionaler Erwerbslosigkeit mittels Mobilfunkdaten für funktionale Stadtgebiete in Deutschland

In mehreren Ländern führt das anhaltende Wachstum der Städte mit ihren besseren Beschäftigungsmöglichkeiten zu verstärkten arbeitsbedingten Pendlerströmen. Obwohl immer mehr Menschen in die Städte pendeln und ziehen, weist der Arbeitsmarkt in städtischen Gebieten höhere Erwerbslosenquoten auf als das Umland. Dieses Phänomen wird auf regionaler Ebene untersucht, und dabei eine alternative Definition der Erwerbslosenquote verwendet, die das Pendlerverhalten einbezieht. Für das deutsche Bundesland Nordrhein-Westfalen werden Daten aus der Arbeitskräfteerhebung mit dynamischen Mobilfunkdaten unter Verwendung von Small-Area-Modellen kombiniert. Aus methodischer Sicht wird ein transformiertes Fay-Herriot-Modell mit Verzerrungs-Korrektur zur Schätzung der Erwerbslosenquoten angewandt. Unter Einbezug der Verzerrungs-Korrektur wird die mittlere quadratische Abweichung mit einem parametrischen Bootstrap-Verfahren geschätzt. Die Leistungsfähigkeit der vorgeschlagenen Methode wird in einer Fallstudie unter Verwendung von amtlichen Daten sowie in modellbasierten Simulationen untersucht. Die Ergebnisse der Anwendung zeigen, dass die (um Pendler bereinigten) Erwerbslosenquoten deutscher Städte niedriger sind als traditionell ermittelte amt-

liche Erwerbslosenquoten indizieren.

Schlüsselwörter: Verzerrungs-Korrektur, Fay-Herriot-Modell, mittlere quadratische Abweichung, Small-Area-Schätzung, Erwerbslosenquoten

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.

Bamberg, July 8, 2022

_____

Nora Würz
July 8, 2022