A Framework for the Estimation of Disaggregated Statistical Indicators Using Tree-Based Machine Learning Methods

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

vorgelegt von
Patrick Krennmair
geboren in Grieskirchen

Berlin, 2022

Patrick Krennmair, *A Framework for the Estimation of Disaggregated Statistical Indicators Using Tree-Based Machine Learning Methods,*
October 2022

Supervisors:
Prof. Dr. Timo Schmid (Freie Universität Berlin) - Erstgutachter
Prof. Nikos Tzavidis, Ph.D. (University of Southampton) - Zweitgutachter

Date of defense:
December 19, 2022

Location:
Berlin

Acknowledgements

First and foremost, I would like to express my deepest respect and gratitude to my supervisor, Prof. Dr. Timo Schmid. He provided the optimal balance between profound statistical guidance and scientific freedom, which was the major key to this thesis.
My sincere thanks addresses Prof. Dr. Nikos Tzavidis, who provided scientific expertise and intuition. His invitation to Southampton for fruitful discussions had a major impact on the realization of this thesis and my personal development as a researcher.
I am thankful for the enjoyable working environment at the Chair of Applied Statistics at Freie Universität Berlin and the Statistical Consulting Unit *fu:stat*. Alejandra, Felix, Marina, Nicolas, Noah, Nora, Sylvia, and Sören were colleagues and became friends. I also want to thank Angelika Wnuk for her kind and reliable administrative support. Without saying, goes my appreciation for Prof. Dr. Ulrich Rendtel, who is an inspiring example of a passionate and dedicated researcher far beyond his statistical expertise. Additionally, I want to thank Dr. Ann-Kristin Kreutzmann for her constructive criticism and friendship.
I gratefully acknowledge the prior support in Vienna throughout my graduate studies. I want to thank Prof. Dr. Alexander K. Wagner, who inspired me to become an independent researcher and who remained a reliable academic mentor. Additionally, I am indebted to Prof. Dr. Karl Milford, who taught me how to maintain a critical perspective and most importantly, how to formulate it scientifically.
My path of education was complemented by so many friendships for which I am grateful. Especially, I want to thank Dominik Bumberger, Martin Thalhammer, Georg Hohensinner, Lena Hebenstreit, Marie-Christin Hiesberger, Lukas Lenz, Jakob Schlösinger, and so many more just for being themselves. A special thanks goes to Thomas Vanek for his cheerful moral support.

I dedicate this work to my family. This includes my grandparents (Hilda & Siegfried and Maria & Alois), who, although encountering unequal educational opportunities, permanently highlighted the importance of knowledge and science. My deepest gratitude is owed to my loving parents Margit und Erwin and my brilliant sister Angelika for their unconditional support, belief, and patience. No words on paper could ever describe how much you mean to me. In the same spirit, I must thank my partner and love Sina Löw for her everlasting and unconditional encouragement.

Publication List

The publications listed below are the result of the research carried out in this thesis titled, "A Framework for the Estimation of Disaggregated Statistical Indicators Using Tree-Based Machine Learning Methods."

1. Krennmair, P. and Schmid, T. (2022) **Flexible domain prediction using mixed effects random forests**, Journal of the Royal Statistical Society: Series C (Applied Statistics) 71(5), 1865-1894. Available from: `https://doi.org/10.1111/rssc.12600`. Accepted and published.

2. 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 from: `https://arxiv.org/abs/2204.10736`.

3. Krennmair, P., Schmid, T., and Tzavidis, N. (2022) **The estimation of poverty indicators using mixed effects random forests: case study for the Mexican state of Veracruz.** *Working paper*, to be submitted.

4. Krennmair, P. (2022) **The R package SAEforest**, *R package vignette.* Available from: `https://CRAN.R-project.org/package=SAEforest`. Accepted and published.

Contents

Introduction

Global and multidimensional problems, such as the anthropogenic climate change or the conditions of persistent inequality demand permanent political and economic awareness. Sustainable and progressive solution strategies presuppose reliable empirical evidence to plan, monitor and adjust policy measures (Lu et al., 2015; Sachs et al., 2019). The sustainable development goals (SDG) are 17 time-bound goals by the United Nations to motivate strategies boosting global equality and prosperity (United Nations, 2015). In order to provide solutions to the major global challenges of the 21st century, multilateral organizations and development agencies acknowledge the importance of increasing efforts to prepare and use data (United Nations, 2014; World Bank Group, 2015; Asian Development Bank, 2021).

Available national-level statistical indicators are often inadequate to detect vulnerability of distinct geographical or demographic groups at desired levels of precision. The identification and elimination of structural patterns of inequality coincides with the fundamental principle of 'leaving no one behind', making systematic data disaggregation a distinct SDG target (SDG-target 17.18) (United Nations, 2015). From a methodological perspective, the dual goal of achieving detailed and reliable estimates from national sample surveys on highly disaggregated 'areas' or 'domains' (e.g. geographical or demographic groups) is referred to as Small Area Estimation (SAE) (Pfeffermann, 2013; Rao and Molina, 2015; Tzavidis et al., 2018).

SAE combines research on statistical procedures to obtain efficient and precise estimates for (non-linear) economic and inequality estimators on hierarchically disjoint levels. Depending on the intended degree of precision, domain-specific sample sizes become small or even zero. Resolving the problem of unreliable estimates due to small sample sizes, model-based approaches link existing auxiliary information from administrative data sources (e.g. census data) or alternative data sources (e.g. mobile phone or remote sensing data) using predictive models (Marchetti et al., 2015; Schmid et al., 2017; Tzavidis et al., 2018; Wardrop et al., 2018). In the existing literature, linear mixed models (LMMs) are predominantly used for prediction tasks, because LMMs control for domain-specific dependencies using random effects.

This thesis introduces and discusses the use of flexible and adaptive algorithms for the predictive purpose of SAE. Particularly, I focus on random forests, which combine individual nonparametric regression trees through bootstrap aggregation (Breiman, 1996, 2001a) and exhibit excellent predictive performance under less restrictive assumptions compared to linear models. Random forests automatically detect and exploit (high-order) relations between predictive covariates, which reduces the risk of model-misspecification for independent unit-level data. Mixed effects random forests (MERFs) (Hajjem et al., 2014) combine the ability to control for dependency structures of survey data with the advantages of regression forests (e.g.

robustness against outliers and implicit model-selection). This thesis aims to bridge concepts by remaining within the general paradigm of SAE, while simultaneously highlighting and evaluating the potential of predictive algorithms. In a broader sense, this includes a methodological commitment to basic concepts of statistical inference and official statistics by accounting for survey-specific dependency structures.

The thesis combines four papers that introduce a coherent framework based on MERFs for the estimation of spatially disaggregated economic and inequality indicators and associated uncertainties. Chapter 1 focusses on flexible domain prediction using MERFs. We discuss characteristics of semi-parametric point and uncertainty estimates for domain-specific means. Extensive model- and design-based simulations highlight advantages of MERFs in comparison to 'traditional' LMM-based SAE methods. Chapter 2 introduces the use of MERFs under limited covariate information. The access to population-level micro-data for auxiliary information imposes barriers for researchers and practitioners. We introduce an approach that adaptively incorporates aggregated auxiliary information using calibration-weights in the absence of unit-level auxiliary data. We apply the proposed method to German survey data and use aggregated covariate census information from the same year to estimate the average opportunity cost of care work for 96 planning regions in Germany. In Chapter 3, we discuss the estimation of non-linear poverty and inequality indicators. Our proposed method allows to estimate domain-specific cumulative distribution functions from which desired (non-linear) poverty estimators can be obtained. We evaluate proposed point and uncertainty estimators in a design-based simulation and focus on a case study uncovering spatial patterns of poverty for the Mexican state of Veracruz. Additionally, Chapter 3 informs a methodological discussion on differences and advantages between the use of predictive algorithms and (linear) statistical models in the context of SAE. The final Chapter 4 complements the previous research by implementing discussed methods for point and uncertainty estimates in the open-source R package **SAEforest**. The package facilitates the use of discussed methods and accessibly adds MERFs to the existing toolbox for SAE and official statistics.

Overall, this work aims to synergize aspects from two statistical spheres (e.g. 'traditional' parametric models and nonparametric predictive algorithms) by critically discussing and adapting tree-based methods for applications in SAE. In this perspective, the thesis contributes to the existing literature along three dimensions: 1) The methodological development of alternative semi-parametric methods for the estimation of non-linear domain-specific indicators and means under unit-level and aggregated auxiliary covariates. 2) The proposition of a general framework that enables further discussions between 'traditional' and algorithmic approaches for SAE as well as an extensive comparison between LMM-based methods and MERFs in applications and several model and design-based simulations. 3) The provision of an open-source software package to facilitate the usability of methods and thus making MERFs and general SAE methodology accessible for tailored research applications of statistical, institutional and political practitioners.

Chapter 1

Flexible domain prediction using mixed effects random forests

Chapter 2

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

2.1 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, 2001a) 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 2.2.1 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 2.2.2 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 2.2.3. As a result, we propose a best practice strategy to ensure the proper usability of EL calibration weights in the context of SAE. Section 2.3 introduces a non-parametric bootstrap-scheme for the estimation of the area-level MSE. In Section 2.4, 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 2.5, 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 2.5.1, 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 2.6, we conclude and motivate further research.

2.2 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 2.2.2 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.

2.2.1 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:

$$y_{ij} = f(\mathbf{x}_{ij}) + u_i + e_{ij} \quad \text{with} \quad u_i \sim N(0, \sigma_u^2) \quad \text{and} \quad e_{ij} \sim N(0, \sigma_e^2). \tag{2.1}$$

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, 2001a) 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 (2.1) 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, 2001a) 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 (2.1) (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, 2001a; 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:

$$\hat{\mu}_i^{\text{MERF}} = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \hat{u}_i = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i} \left(\frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - \hat{f}(\mathbf{x}_{ij})) \right), \tag{2.2}$$

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

2.2.2 MERFs under aggregated data

Estimates for the area-level mean μ_i using MERFs from Equation (2.2) 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 (2.2) 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:

$$\hat{\mu}_i^{\text{MERFagg}} = \sum_{j=1}^{n_i} \hat{w}_{ij} \left[\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i \right]. \tag{2.3}$$

Note that optimal estimates for required model components $\hat{f}$ and $\hat{u}_i$ are obtained similar to Equation (2.2) from survey data using the MERF algorithm as described by Krennmair and Schmid (2022). We incorporate aggregate census-level covariate information through the calibration 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:

$$\hat{w}_{ij} = \frac{1}{n_i} \frac{1}{1 + \hat{\lambda}_i^\intercal (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i})}, \tag{2.4}$$

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

2.2.3 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 (2.4). 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 (2.4). 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 like-

lihood 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 2.4.

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 (2.4) 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 (2.4) 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 $\hat{\mu}_i^{\text{MERFagg}}$ (2.3) including weights $\hat{w}_{ij}$ following Equation (2.4).

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 (2.4).

(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 "similar" 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 (2.3).

The general performance is illustrated by the results of the model-based simulation in Section 2.4. Furthermore, the proposed best-practice strategy will be demonstrated in the application in Section 2.5.

2.3 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. (2021) 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 (2.3). 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 (2.1). 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 (2.2), $\hat{\sigma}_\epsilon^2$ 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 (2.3) 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:

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

(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 (2.3) 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 $i = 1, ..., D$

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

2.4 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 2.2 and Section 2.3 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*, (2.3)) 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*, (2.2)) 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*).

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 2.5 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 2.1. 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.

Table 2.1: 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 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 2.3.

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.

2.4.1 Performance of point estimators of the small area means

We start with a focus on the performance of point estimates. Figure 2.1 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 2.2.3 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 (2.4) 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

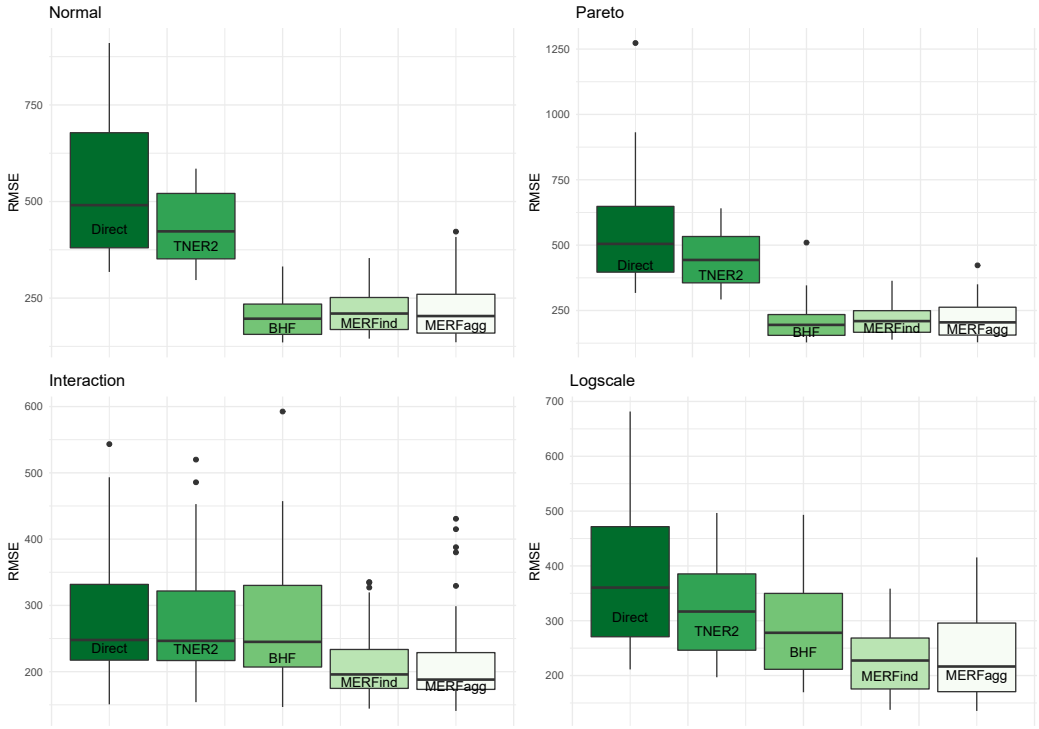


Figure 2.1: Empirical RMSE comparison of point estimates for area-level averages under four scenarios

for the calculation of calibration weights.

In the *Interaction* scenario, the point estimates of the proposed *MERFagg* outperform traditional 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 2.1 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

Table 2.2: 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

of survey data for a respective area i as discussed in Section 2.2.2.

Table 2.2 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 2.1), the median and mean values of *MERFagg* are remarkably low emphasizing the quality of estimates given the the substantial reduction in required covariate information.

2.4.2 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 2.3, 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 2.2 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 2.3 is mainly driven by overestimation in areas with low sample

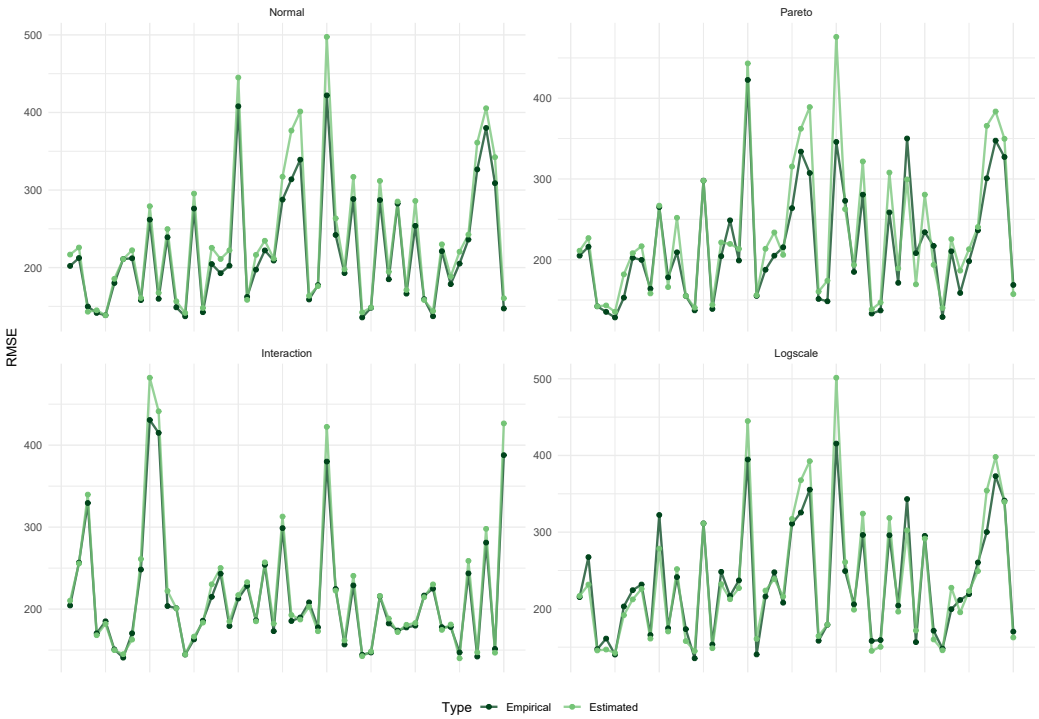


Figure 2.2: Estimated and empirical area-level RMSEs for four scenarios

Table 2.3: 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

sizes. Thus, our non-parametric MSE estimator provides an upper bound for the uncertainty of 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.

2.5 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).

2.5.1 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.

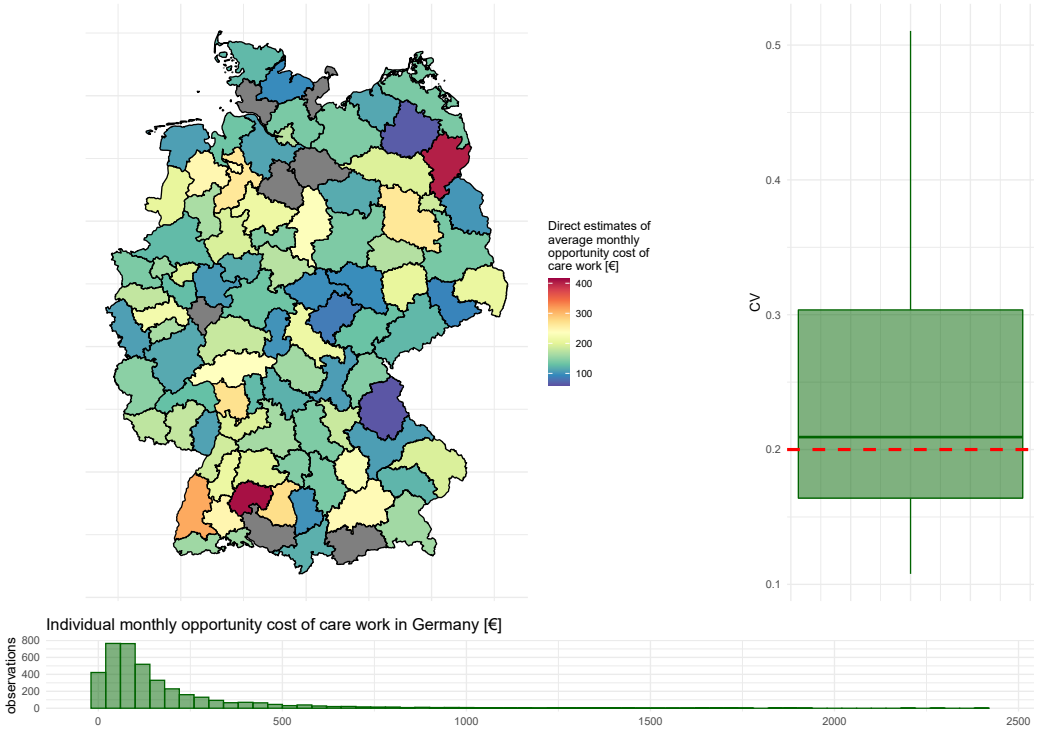


Figure 2.3: Overview of *direct* estimates, corresponding CVs and the distribution of opportunity cost of care work in Germany.

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 2.3.

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 (2.3). 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 (2019) postulates that estimates with a CV of less than 20% can be considered as reliable. As reported by Figure 2.3, more than half of the regions (47 out-of remaining 89) exceed this threshold.

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.

2.5.2 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 2.3 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 2.3, 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

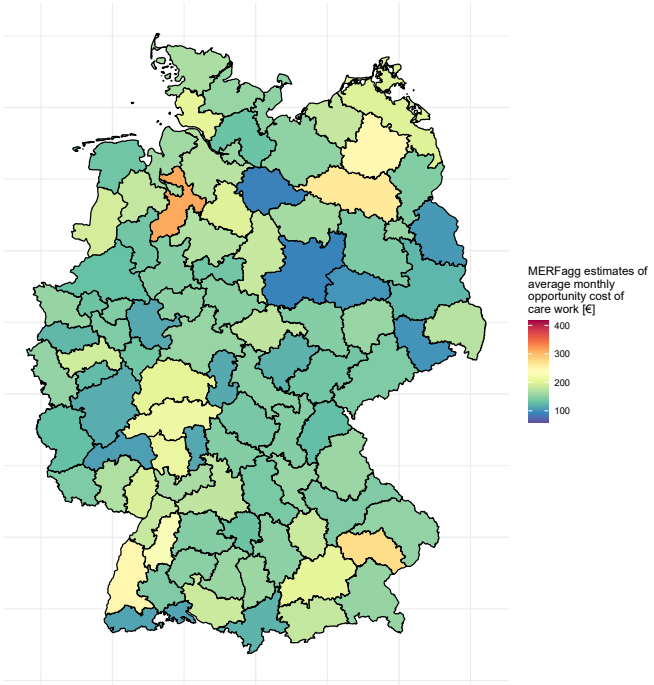


Figure 2.4: Spatial representation of area-level mean estimates from *MERFagg* (2.3) for mean monthly opportunity cost of care work [€].

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 2.6. 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 (2.3) 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 2.3 and estimates based

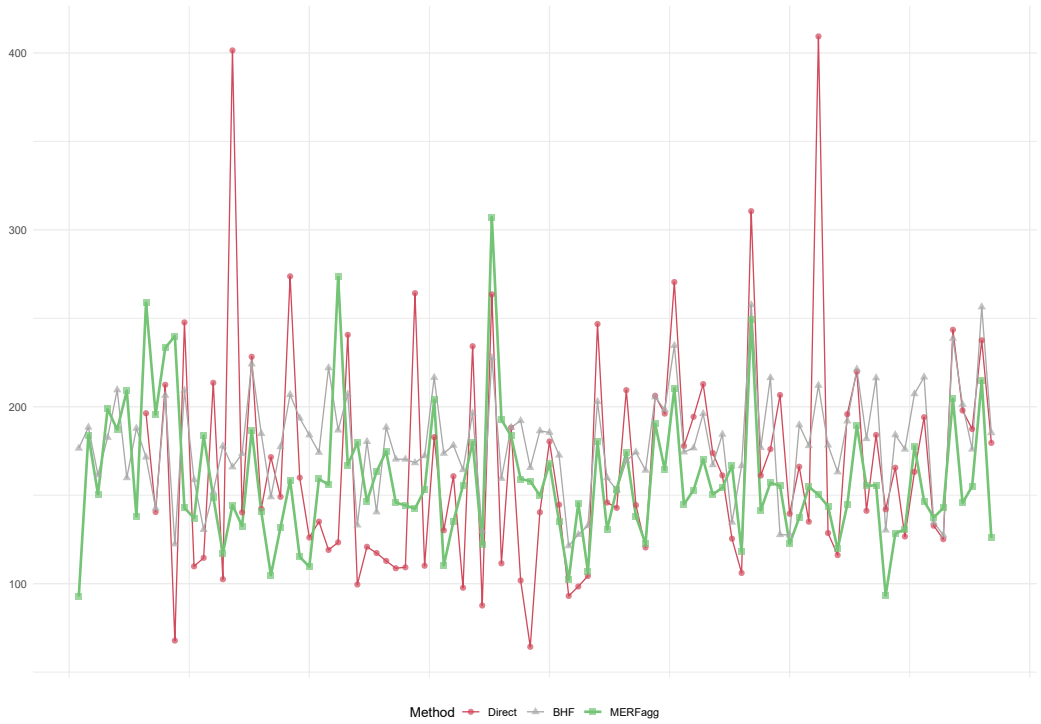


Figure 2.5: Detailed comparison of area-level mean estimates for monthly opportunity cost of care work [€]. The 96 German RPRs are sorted by increasing sample size. We compare results based on methods *direct*, *BHF*, and *MERFagg*.

on *MERFagg* from Figure 2.4 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 2.5 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 2.6 juxtaposes CVs for *direct* estimates, the *BHF*, and our proposed method of *MERFagg* to contextualize the performance of point estimates from Figure 2.5. 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 2.6, 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

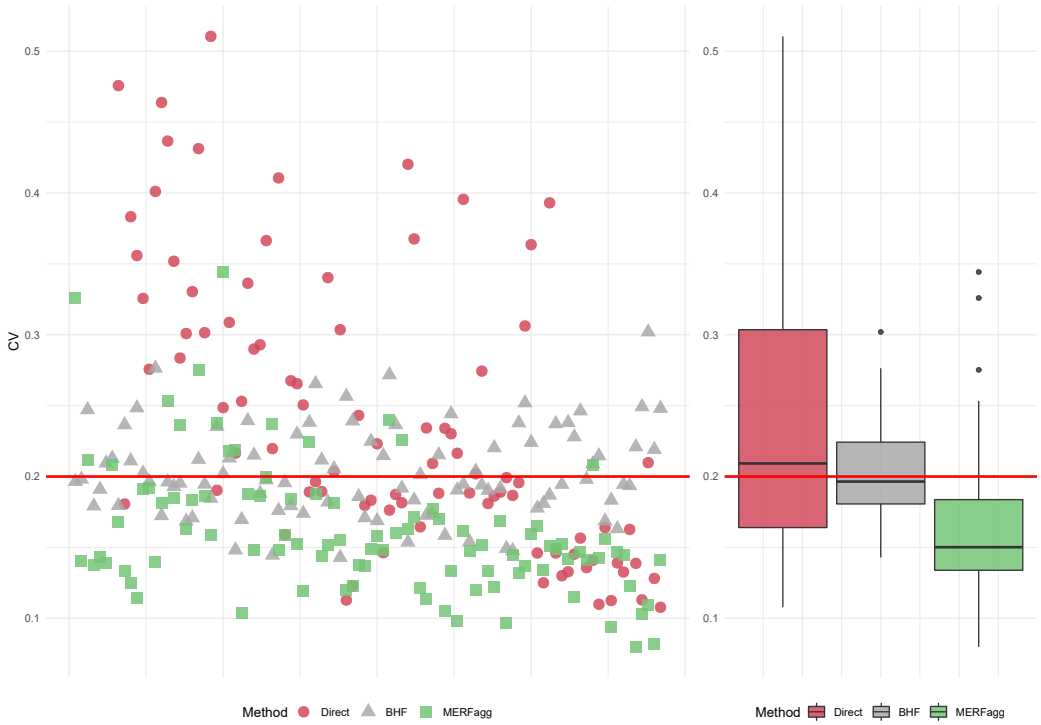


Figure 2.6: Left: Comparison of area-specific CVs ordered from low to high sample sizes. Right: Comparison of CVs over 96 respective areas between direct, *BHF* and *MERFagg*. The red line marks the 20%-criterion for defining reliable estimates by Eurostat (2019).

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 2.4 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.

2.6 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 (2.1) allows for f to adapt any functional form regarding the estimation of the conditional mean of y_{ij} given $\mathbf{x}_{ij}$ 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 2.5)

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

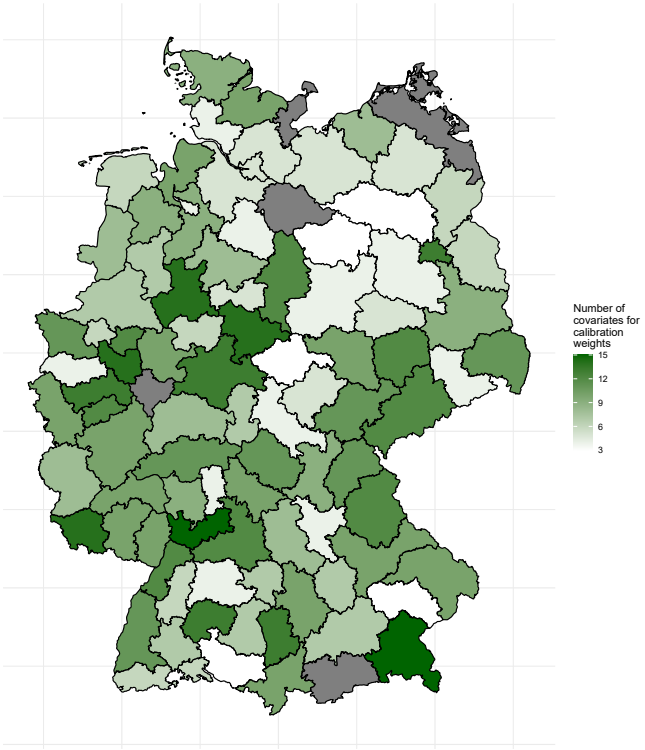


Figure B.1: Inclusion of covariates through weights

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 (2.3) 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:

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

Chapter 3

The estimation of poverty indicators using mixed effects random forests: case study for the Mexican state of Veracruz

3.1 Introduction

The sustainable development goals (SDGs) are a committed agenda by the United Nations, comprising 169 targets for 17 time-bound goals on equality and prosperity under the general premise of 'leaving no one behind' (United Nations, 2015). SDG1 addresses the eradication of poverty. Knowledge on the deprivation at low geographic or administrative levels (e.g. 'areas' or 'domains') facilitates efficient allocation of aid. Data gaps constrain reliable disaggregation of indicators to monitor SDGs on a finer spatial scale. Available national statistical indicators are inadequate to monitor tailored policies and potentially hide vulnerable groups in national-level aggregates (Wardrop et al., 2018). From a methodological perspective, this dual goal of achieving detailed and reliable estimates from national sample surveys on highly disaggregated geographical and other domains (e.g. demographic groups) is scientifically referred to as Small Area Estimation (SAE) (Pfeffermann, 2013; Rao and Molina, 2015; Tzavidis et al., 2018).

Direct domain-specific estimates of statistical indicators are prone to undermine tolerable levels of reliability due to the inverse relation between the level of disaggregation and correspondingly decreasing sample sizes on the scale of interest. Multilateral organizations increasingly acknowledge the use of model-based SAE to close existing data gaps by combining official (or alternative) data sources with survey data and subsequently provide reliable and precise area-level estimates for desired linear and non-linear indicators (World Bank Group, 2015; Asian Development Bank, 2021). Kilic et al. (2017) maintain that model-based SAE significantly mitigates survey costs for multilateral organizations, while ensuring reliability of estimates. Reductions of multiple inequalities between and within countries depend on the formulation of transformative strategies based on reliably measurable metrics combing profound research and data-sources (Lu et al., 2015; Sachs et al., 2019). In addition, the fundamental

principle of 'leaving no one behind' is inevitably connected to disaggregated measures monitoring progress for demographic or geographic subgroups (Asian Development Bank, 2021). In fact, systematic data disaggregation is a distinct SDG target (SDG-target 17.18) (United Nations, 2015).

In this paper, we propose a flexible, data-driven, and semi-parametric alternative for the estimation of non-linear domain-specific indicators using mixed effects random forests (MERF). Random forests (Breiman, 2001a) excel in terms of predictive performance without explicit model assumptions in the presence of skewed data and outliers and are applicable to high-dimensional data (Hastie et al., 2009; Varian, 2014). Few tuning parameters and automated model-selection including the detection of complex and higher order interactions of covariates justify their popularity as non-parametric prediction algorithms (Biau and Scornet, 2016). MERFs (Hajjem et al., 2014) are a composite model linking random forests for fixed effects with a structural component, accounting for hierarchical dependencies of survey data with random effects. Krennmair and Schmid (2022) introduce MERFs in the methodological tradition of SAE to estimate area-level means and extensively evaluate the performance of point and uncertainty estimators under design- and model-based simulations. The major methodological contribution of this paper is the extension of the approach by Krennmair and Schmid (2022) towards an estimator for a finite-population cumulative distribution function (CDF) from survey sample data as originally proposed by Chambers and Dunstan (1986) (CD). Resulting estimates of domain-specific CDFs directly allow to derive (non-linear) indicators. Additionally, we propose two non-parametric MSE bootstrap schemes to assess the uncertainty of domain-specific estimates.

A distinct advantage of our proposed approach is the generic robustness against model-failure (e.g. flexible protection against model-misspecification, valid variable selection and the effective handling of outliers) (Jiang and Rao, 2020). 'Traditional' unit-level SAE methods, which combine data sources to estimate non-linear indicators, are regression-based and rely predominantly on the theoretical framework of linear mixed models (LMM) (Rao and Molina, 2015). LMMs rely on Gaussian assumptions that hardly meet empirical evidence for economic and inequality data. Well known examples for poverty mapping are the empirical best predictor (EBP) (Molina and Rao, 2010) or the World Bank Method (ELL) proposed by Elbers et al. (2003). Several strategies evolved to counter improperly met assumptions and associated bias in point and uncertainty estimates by advanced transformation strategies on the dependent variable (Sugasawa and Kubokawa, 2017; Tzavidis et al., 2018; Sugasawa and Kubokawa, 2019; Rojas-Perilla et al., 2020). Another strategy to mitigate effects of failed model assumptions, is the formulation of predictive models under more flexible distributions (Diallo and Rao, 2018; Graf et al., 2019). Alternatively, Tzavidis et al. (2018) propose a method based on M-quantile models, which are a robust method avoiding distributional assumptions including the formal specification or area-level random effects. Recently, Marchetti and Tzavidis (2021) scrutinize the estimation of inequality indicators using M-quantile models.

Despite conceptual differences between machine learning and 'traditional' statistical methods (e.g. best possible predictions vs. parametric representation and interpretation), machine learning methods became a substantial element in statistical methodology research (Efron,

2020). In SAE, relatively few studies concern the integration of predictive algorithms. For instance, the comparison of LMM-based and tree-based estimates for sub-populations was investigated by Anderson et al. (2014) in the context of population densities and by De Moliner and Goga (2018) in the context of electricity consumption. Singleton et al. (2020) combine unsupervised learning and boosted regression trees to map digital inequality. Mendez (2008) provides initial theoretical and empirical considerations using random forests for SAE and Dagdoug et al. (2021) analyse theoretical properties of random forests in the context of complex survey data. Bilton et al. (2017) use classification trees to estimate household poverty and Bilton et al. (2020) use regression trees to estimate non-linear poverty indicators.

Apart from its statistical contribution, this paper aims to inform a methodological discussion. Efron (2020) reviews objectives of prediction, estimation and attribution and thereby extends the discourse on two cultures of predictive algorithms and statistical methods initiated by Breiman (2001b). Owed to its purpose, SAE combines concepts of prediction and estimation. We utilize models and auxiliary information from census data for pure prediction and estimate indicators combining observed and predicted values requiring concepts of inference and survey statistics. Using random forests for the primary prediction part is an agnostic option, however, we must meet basic premises and requirements of SAE including a framework for valid inferences and considerations of dependency structures of survey data. From our perspective, the introduction of MERFs for SAE aligns with the postulate of Efron (2020), maintaining that an opportunity for modern statistics lies in critical attempts to make predictive algorithms 'scientifically applicable' to meet methodological requirements of specific applications or statistical subdisciplines.

This paper introduces our proposed method of MERFs to reliably uncover spatial concentrations of poverty measured by the head count ratio (HCR) and the poverty gap (PGAP). Our case study targets economic vulnerability in the Mexican State of Veracruz that might otherwise be hidden in macro-level aggregates. We describe provided data in detail in Section 3.2. The rest of the paper is organized as follows: Section 3.3.1 introduces a general unit-level model framework for MERFs in SAE. Section 3.3.2 focuses on the estimation of area-level CDFs and Section 3.3.3 demonstrates the determination of (non-linear) indicators. Uncertainty estimation based on two elaborate non-parametric bootstrap schemes is the focus of Section 3.3.4. Section 3.3.5 critically reflects on potential improvements concerning the modelling of dependency structures or the introduction of transformation strategies. We assess the quality of point and uncertainty estimates in the design-based simulation in Section 3.4. We discuss results of the application of municipality-level poverty mapping in Veracruz in Section 3.5 and Section 3.6 concludes.

3.2 Veracruz case study: data sources and initial analysis

Monitoring regional aspects of poverty is imperative for progressive policies (such as the SDGs) to translate into inclusive and sustainable actions. Monetary transfer programs in Mexico have been beyond their expectations concerning reductions of multidimensional poverty and inequality (Lambert and Park, 2019). The analysis on sub-national health related SDG

indicators in Mexico exposes spatial heterogeneity and emphasizes the need for tailored policy programs reducing local concentration of poverty to reach the SDG commitments until 2030 (Gutierrez et al., 2020). In this paper, we focus on district-level differences in monetary poverty for one of the 32 federal entities in Mexico. Veracruz is located in the east of the country and is characterized by its long coast with the Gulf of Mexico. According to the sub-national Human Development Index (Smits and Permanyer, 2019), Veracruz is among the least developed states of Mexico. A major characteristic of the state is its geographical, ethnic, and linguistic diversity which transfers into various community structures and economic as well as agricultural systems (Dietz, 2012).

We use data from 2010 provided by CONEVAL (Consejo Nacional de Evaluación de la Polìtica des Desarrollo Social), which combines the Mexican household income and expenditures survey ENIGH (Encuesta Nacional de Ingreso y Gastos de los Hogares) with a sample of census microdata by the National Inistitute of Statistics and Geography (Instituto Nacional de Estadística y Geografía). Our dataset comprises income and socio-demographic data with equally measured variables in the survey as well as the census data. We construct poverty indicators based on the total household per capita income (*ictpc*, measured in pesos), which is exclusively available in the survey. Veracruz is organized into 212 municipalities. The survey data comprises information on 1453 households from 58 municipalities, resulting in domain-specific samples sizes ranging from a minimum of 2 to a maximum of 120 with a median of 19 households. This leaves 154 municipalities out-of-sample. In the census dataset, we have 246899 households from all 212 municipalities. A summary of domain-specific data is provided in Table 3.1.

Table 3.1: Summary statistics on in- and out-of-sample areas: area-specific sample size of census and survey data

	Total 212		In-sample 58		Out-of-sample 154	
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Survey domain sizes	2	10	19	25	33	120
Census domain sizes	393	794	969	1165	1131	7734

Direct domain-level estimates of poverty indicators for Veracruz are possible for 58 out of 212 domains. Even for sampled municipalities, the low domain-specific sample sizes compromise the reliability of point estimates. Statistical offices assess the quality of estimates by its coefficient of variation (CV), which is defined as the indicator's standard deviation relative to its value (Eurostat, 2004). The CVs for direct estimates report high variation with mean/-median values for HCR of 0.415/0.338 and for the PGAP of 0.489/0.413. Direct estimates are mapped in Figure 3.3 in Section 3.5, where we will focus on a detailed interpretation of results. The initial analysis on direct estimates highlights limits concerning the depiction of the spatial distribution of poverty and the identification of geographical hotspots. The availability of equally measured auxiliary data in the census and survey data showcases advantages of model-based SAE methods. An incorporation of covariate census data enables estimates for out-of-sample domains and simultaneously improves the quality of point and uncertainty estimates for sampled areas (Tzavidis et al., 2018).

Predominant unit-level models for the estimation of domain-specific non-linear poverty indicators rely on LMMs (Rao and Molina, 2015). This paper's approach based on MERFs for the estimation of non-linear indicators introduces a new method and simultaneously new perspectives on best practices. We motivate the use of alternative models for the prediction task by suggesting an initial experiment. We divide the survey data 100 times randomly into a 80% training and 20% test set and compare the predictive performance of trained unit-level models for the random forest (RF), the MERF, a linear model (LM) and a LMM. For the computational realization throughout the paper, we use R (R Core Team, 2022). For the initial calculations, we use packages **lme4** (Bates et al., 2015) and **ranger** (Wright and Ziegler, 2017). We summarize the performance in terms of the average mean and median root squared prediction errors in Table 3.2. This repeated prediction experiment is thought to provide initial insights concerning two paradigms of SAE, i.e. predictive advantages of algorithms compared to traditionally used LMMs and the importance of modelling domain-specific variation with random intercepts.

Table 3.2: Comparison of unit-level mean and median root squared prediction error on 100 randomly generated training (80%) and test (20%) set splits on the survey data.

	Median	Mean
MERF	1670.025	677.0150
RF	1718.750	701.5135
LMM	1734.220	794.3899
LM	1787.230	791.7486

Table 3.2 reports increased predictive performance of tree-based methods (RF and MERF) compared to the linear alternatives of LM and LMM. Additionally, models that account for the structural nature of the survey data using random intercepts (MERFs and LMMs) outperform competitors, which neglect dependency structures. Although Table 3.2 approves the use of MERFs, we abstain from over-interpreting the results as the superior performance in this experiment does not automatically transfer to the quality of domain-specific poverty indicators. We compare the performance of 'traditional' SAE methods and MERFs in detail in Section 3.4. Before we proceed with the technical introduction of our proposed method and its subsequent evaluation, we aim to discuss an alternative workflow and diagnostics associated to non-parametric modelling using (ME)RFs.

Considering SAE based on LMMs (e.g. the EBP), the next step towards improved estimates for indicators is achieved by a critically inquiry into the validity of model assumptions and the subsequent use of suitable transformations and procedures of model-selection (Tzavidis et al., 2018). Rojas-Perilla et al. (2020) use the Bayesian Information Criterion (BIC) to identify an optimal model for predicting *ictpc* based on available covariates. In contrast, RFs perform an implicit model-selection (Breiman, 2001a). Although we postpone the full comparison of results based on 'traditional' and tree-based estimates and complementary model properties to Section 3.5, we discuss visual model-diagnostics of random forests. The following figures further motivate the use of tree-based models for SAE and mitigate the general argument that improved prediction serves naturally at the cost of interpretability.

Figure 3.1 demonstrates partial dependence plots (pdp) (Greenwell, 2017) and variable im-

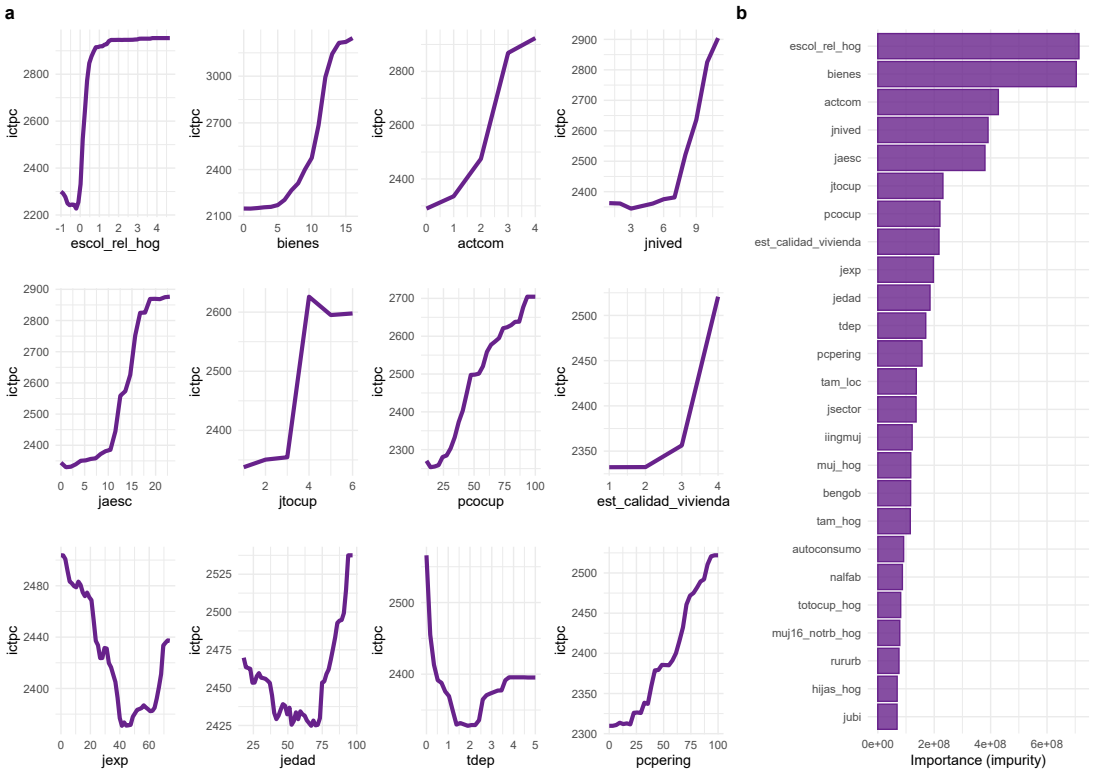


Figure 3.1: Visual diagnostics on predictive relations between dependent variable *ictpc* and predictors. a) Partial dependency plots; b) variable importance plot

portance plots (Greenwell et al., 2020). The pdp plot estimates the marginal effect for specified predictors on the target variable. The variable importance ranks the importance of predictors based on the mean decrease in impurity (variance) calculated for each predictor as the sum over the number of splits across all trees that include the predictors. From the combined information of the two plots, we infer the predictive impact of variables and whether relationships between *ictpc* and corresponding predictors tend to be complex or linear. Figure 3.1b shows the 25 (out of 39) most influential covariates and we observe high rankings for information on education (*escol_rel_hog, jnived, jeasc*), the condition of the household and its goods (*bienes, actcom, jtocup, est_calidad_vivienda*), work-related variables (*pcocup, jexp, pcpering*), and the age (*jedad, tdep*). Table C.1 in the Appendix summarizes details on the variables and additionally states whether the most influential variables are also selected into the optimal EBPbc model. Overall, Figure 3.1 indicates non-linear relations between the target variable and its predictors, such as for *escol_rel_hog, jtocup, jexp, jedad* or *tdep*. This observation promotes the use of methods that implicitly detect and handle complex interactions.

3.3 General model and estimation of finite population parameters

This section introduces the estimation of non-linear indicators within a general semi-parametric framework focusing on MERFs. The construction of relevant area-level metrics presupposes methodology on the estimation of area-level CDF functions for continuous target variables in

the context of SAE. We follow the approach of Chambers and Dunstan (1986) for the estimation of a finite-population CDF of y_{ij} from which non-linear indicators can be directly obtained.

3.3.1 Unit-level models for Small Area Estimation

We assume a finite population P with D disjunct areas P_i of sub-population sizes N_i, where $i = 1, ..., D$ specifies the areas and $N = \sum_{i=1}^{D} N_i$ defines the population size of all areas. The sample s consists of area-specific sub-samples s_i with overall size $n = \sum_{1}^{D} n_i$. In contrast, non-sampled observations are denoted as r_i with size $N_i - n_i$. We denote individual units within each area as $j \in s_i$ for sampled and $j \in r_i$ for unsampled observations. The continuous unit-level target variable is given by y_{ij}, assuming information on y for n observations of our sample. Vector $\mathbf{x}_{ij} = [x_1, x_2, ..., x_p]^\intercal$ captures p auxiliary covariates and auxiliary variables are known for N units in our population P. We assume that y_{ij} follows the semi-parametric general model:

$$y_{ij} = f(\mathbf{x}_{ij}) + v_i + \epsilon_{ij}, \tag{3.1}$$

where

$$\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2) \quad \text{and} \quad v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2).$$

Essentially, Model 3.1 consists of two major parts: the (non-parametric) fixed part of $\mathbf{x}_{ij}$ and the structural part v_i, representing area specific random intercepts, which characterizes small area differences in the conditional distribution of y_{ij} given $\mathbf{x}_{ij}$. Individual independently distributed unit-level errors are denoted by ϵ_{ij}, which are mutually independent to the random effects v_i. Assumed correlations arise only due to between-area variations with associated variance components σ_v^2 (area-specific effects) and σ_ϵ^2 (unit-level effects). Model 3.1 is extendable to capture complex correlation structures as well as higher-order hierarchical dependencies through modifications of respective variance-covariance matrices of random components. A full discussion is beyond the scope of this paper, however, we motivate a general formulation of Model 3.1 in Krennmair and Schmid (2022).

The goal is the estimation of a relation f between covariates $\mathbf{x}_{ij}$ and the target variable y_{ij} using a survey sample, to provide values for non-sampled observations of y_{ij} utilizing available supplementary covariate data and information across areas. In general, f can be any parametric or non-parametric function expressing the conditional mean of target variable y given $\mathbf{x}$. The general formulation of Model 3.1 synergizes several unit-level SAE-models. For instance, the linear nested-error unit-level model proposed by Battese et al. (1988) is a special case of Model 3.1. We state f to be the linear model $f(\mathbf{x}_{ij}) = \mathbf{x}_{ij}^\intercal \beta$, with regression parameters $\beta = [\beta_1, ..., \beta_p]$. This setting also defines the structural baseline for the simulation of the conditional distribution in the case of the EBP (Molina and Rao, 2010) as well as the EBP under data-driven transformation (Rojas-Perilla et al., 2020). Defining f in Model 3.1 as a random forest, results in the MERF-approach proposed by Hajjem et al. (2014), which is the preferred specification throughout the rest of the paper.

Model 3.1 expresses the conditional mean of a continuous unit-level dependent variable. While Section 3.3.2 focusses on the estimation of the area-specific CDF from which (non-) linear indicators are obtainable, we scrutinize optimality estimates for the model parameters

$f, \sigma_v^2, \sigma_\epsilon^2$ first. For the linear nested-error unit-level model, optimal parameters are found by ML or REML (Battese et al., 1988; Rao and Molina, 2015). For fitting Model 3.1, where f is a random forest, we use an approach reminiscent of the EM algorithm similar to Hajjem et al. (2014). In short, 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, 2001a; Biau and Scornet, 2016). The proposed algorithm is as follows:

1. Initialize $b = 0$ and set random components $\hat{v}_{(0)}$ to zero.
2. Set $b = b + 1$. Update $\hat{f}(\mathbf{x}_{ij})_{(b)}$ and $\hat{v}_{(b)}$:
   (a) $y^*_{ij,(b)} = y_{ij} - \hat{v}_{i,(b-1)}$
   (b) Estimate $\hat{f}()_{(b)}$ using a random forest with dependent variable $y^*_{ij,(b)}$ and covariates $\mathbf{x}_{ij}$. Note that $\hat{f}()_{(b)}$ is the same function for all areas i.
   (c) Get the OOB-predictions $\hat{f}(\mathbf{x}_{ij})^{OOB}_{(b)}$.
   (d) Fit a linear mixed model without intercept and restricted regression coefficient of 1 for $\hat{f}(\mathbf{x}_{ij})^{OOB}_{(b)}$:
   $$y_{ij} = \hat{f}(\mathbf{x}_{ij})^{\text{OOB}}_{(b)} + \hat{v}_{i,(b)} + \epsilon_{ij}.$$
   (e) Extract the estimated variance components $\hat{\sigma}^2_{\epsilon,(b)}$ and $\hat{\sigma}^2_{v,(b)}$ and estimated random effects $\hat{v}_{(b)}$. Note that the random effect for area i is calculated as:
   $$\hat{v}_{i,(b)} = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_\epsilon^2/n_i} \left(\frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - \hat{f}^{\text{OOB}}_{(b)}(\mathbf{x}_{ij})) \right)$$
3. Repeat Step (2) until convergence is reached.

The convergence of the algorithm is assessed by the marginal change of the modified generalized log-likelihood (GLL) criterion:

$$GLL(f, v_i|y_{ij}) = \sum_{i=1}^{D} ([y_{ij} - f(\mathbf{x}_{ij}) - v_i]^\intercal I_i \sigma_\epsilon^{-1} [y_{ij} - f(\mathbf{x}_{ij}) - v_i] + v_i^\intercal \sigma_v^{-1} v_i + \log|\sigma_v| + \log|I_i \sigma_\epsilon|) \tag{3.2}$$

In the linear case with $f() = \mathbf{x}_{ij}^\intercal \beta$, and for given variance components σ_ϵ and σ_v, the maximization of the GLL-criterion is equivalent to the solution of so-called mixed model equations (Wu and Zhang, 2006), leading to the best linear unbiased predictor (BLUP) for every out-of-sample unit $j \in r_i$ for each area i:

$$\mu_{ij} = f(\mathbf{x}_{ij}) + v_i = \mathbf{x}_{ij}^\intercal \beta + \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\epsilon^2/n_i} \left(\frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - \mathbf{x}_{ij}^\intercal \beta) \right)$$

A similar result holds, when we assign f to be a random forest: the corresponding solution for the random intercept v_i, optimizing the GLL-criterion 3.2 for known parameters σ_v and σ_ϵ

is given by:

$$v_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\epsilon^2/n_i} \left(\frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - f(\mathbf{x}_{ij})) \right) \tag{3.3}$$

Mathematical details of the derivations are provided in Krennmair and Schmid (2022). After the convergence of the algorithm, we propose $\hat{v}_i$ from Step 2.e) to be a suitable estimator for v_i. This result is in line with Capitaine et al. (2021), claiming that $\hat{v}_i$ under the EM-based algorithm is obtained by taking the conditional expectation given the data y_{ij} and subsequently $\hat{v}_i$ can be considered as the EBLUP for the linear part of Model 3.1. Thus, we propose $\hat{\mu}_{ij}$ as a suitable estimator for individual out-of-sample observations under our proposed general mixed model:

$$\hat{\mu}_{ij} = \hat{f}(\mathbf{x}_{ij}) + \hat{v}_i = \hat{f}(\mathbf{x}_{ij}) + \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_\epsilon^2/n_i} \left(\frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - \hat{f}(\mathbf{x}_{ij})) \right) \tag{3.4}$$

3.3.2 Estimation of finite population parameters

In the context of unit-level mixed models, Krennmair and Schmid (2022) use Estimator 3.4 directly to estimate area-level means. For the estimation of quantiles or (non-linear) poverty indicators, such as the FGT-indicators (Foster et al., 1984), we need information on the area-specific CDF of y_{ij}. Chambers and Dunstan (1986) develop a model-consistent estimator for a finite-population CDF from survey sample data and provide asymptotic results under the linear mixed model. Essentially, Chambers and Dunstan (1986) combine a model for a finite-population CDF of y_{ij} with the smearing-approach by Duan (1983). The concept of smearing relates to the bootstrap principle, as the unknown error distribution is constructed using the empirical CDF of the regression residuals and subsequently expected values of the resulting error distribution are taken. Tzavidis et al. (2010) introduce the use of the CD-method within a general unit-level framework and focus on the estimation of SAE means and quantiles. Under an unit-level estimator expressing the expected value conditional on area i for a non-sampled observation j (such as $\hat{\mu}_{ij}$ in Equation 3.4), Tzavidis et al. (2010) express the CDF-estimator for area i as:

$$\hat{F}_i^{CD}(t) = N_i^{-1} \left[\sum_{j \in s_i} I(y_{ij} \leq t) + n_i^{-1} \sum_{j \in s_i} \sum_{k \in r_i} I\left(\hat{\mu}_{ik} + (y_{ij} - \hat{\mu}_{ij}) \leq t\right) \right] \tag{3.5}$$

Originally, Tzavidis et al. (2010) estimate area-level means and quantiles assuming $\hat{\mu}_{ij} = \mathbf{x}_{ij}^\intercal \hat{\beta}_\Psi(\hat{\theta}_i)$, where $\hat{\beta}_\Psi(\hat{\theta}_i)$ is a robust, M-quantile estimator (Breckling and Chambers, 1988). However, the proposed framework enables the estimation of various social inequality and poverty indicators (Marchetti et al., 2012; Marchetti and Tzavidis, 2021) and simultaneously allows for a broader class of models: for instance, Tzavidis et al. (2010) emphasize that in the context of unit-level linear models, a bias-adjusted alternative to the EBLUP is achieved by substituting $\hat{\mu}_{ij} = \mathbf{x}_{ij}^\intercal \hat{\beta} + \hat{v}_i$.

Based on Model 3.1, we propose a modified estimator of $F_i^*(t)$ for MERFs building on the CD-method from Equation 3.5. We use $\hat{\mu}_{ij} = \hat{f}(\mathbf{x}_{ij}) + \hat{v}_i$, where $\hat{f}$ is a random forest and esti-

mates $\hat{f}$ and $\hat{v}$ are obtained from the MERF-algorithm presented in Section 3.3.1. Additionally, we propose the use of OOB-residuals $e^*_{ij} = y_{ij} - \hat{\mu}^{\text{OOB}}_{ij}$, where $\hat{\mu}^{\text{OOB}}_{ij} = \hat{f}(\mathbf{x}_{ij})^{\text{OOB}} + \hat{v}_i$. Using OOB-residuals is a simple and genuine solution to achieve more robust estimates of the CDF for MERFs. Moreover, we ensure that these model residuals e^*_{ij} mirror the estimated variance properties under our Model 3.1. Our proposed estimator is given by:

$$\hat{F}^*_i(t) = N_i^{-1}\left[\sum_{j\in s_i} I(y_{ij} \leq t) + n_i^{-1}\sum_{j\in s_i}\sum_{k\in r_i} I\left(\hat{\mu}_{ik} + \underbrace{(y_{ij} - \hat{\mu}^{\text{OOB}}_{ij})}_{e^*_{ij}} \leq t\right)\right] \tag{3.6}$$

Estimating $\hat{F}^*_i(t)$ based on a smearing approach is computationally intensive. Monte Carlo (MC) approximations are an alternative to the smearing-type Estimator 3.5 for area-specific CDFs. Marchetti et al. (2012) and Marchetti and Tzavidis (2021) discuss MC-based alternatives in the context of M-quantile models. MC-based alternatives approximate $E(y_{ik}|y_s;\hat{c})$, where $\hat{c}$ is an suitable estimator for c, which captures unknown super-population parameters. This draws conceptual parallels to the EBP. The EBP builds on a firm theoretical background rooted within the methodology of empirical Bayes (EB) for which statistical properties based on structural and distributional assumptions have been scrutinized (Molina and Rao, 2010; Rao and Molina, 2015). To the best of our knowledge, there exist no theoretical considerations extending the EBP towards semi- or non-parametric model classes. A solution to this discussion exceeds the purpose of this paper, but provokes further research. We provide details on an algorithm for an MC approximation to Equation 3.6 in the Appendix. In short, our empirical observations coincide with the propositions of Marchetti et al. (2012) and Marchetti and Tzavidis (2021), stating that no systematic differences in the quality of point estimates for various indicators between the discussed smearing and MC-based approaches is observable.

3.3.3 Estimation of poverty indicators

In this section, we focus on the calculation of selected area-level indicators obtainable from the area-level CDF-Estimator 3.6. Based on $\hat{F}^*_i$, we can directly obtain desired domain-specific estimators following a flexible and convenient strategy: let δ_i be a parameter of interest for area i and $h()$ is a function that calculates the indicators. We can write $\delta_i = h(y_{s_i} \cup y_{r_i})$, where s_i are sampled and r_i are non-sampled observations. The resulting estimate for indicator δ_i is defined as:

$$\hat{\delta}_i = h(y_{s_i} \cup \hat{y}^*_{i(kj)}),$$

where $\hat{y}^*_{i(kj)}$ is the vector of smearing values of length $j \times k$ from Equation 3.6 given by $\hat{y}^*_{i(kj)} = \hat{\mu}_{ik} + e^*_{ij}$.

For instance, the first key indicator for the eradication of poverty (HCR) is a realization of a more general class of poverty indicators referred to as so-called FGT-indicators (Foster et al., 1984):

$$FGT_{ij}(\alpha, t) = \left(\frac{t - y_{ij}}{t}\right)^{\alpha} I(y_{ij} \leq t),$$

where t states a predefined poverty line. Setting $\alpha = 0$ gives the HCR and setting $\alpha = 1$

defines the PGAP. While the HCR simply refers to the proportion of households with income below a defined poverty line t, the PGAP measures poverty intensity by quantifying the degree of average income difference to the poverty line of people below the poverty line relative to the poverty line. For our example assuming $h() = FGT_{ij}(\alpha, t)$, area-specific estimates on required indicators can be obtained as follows:

$$\widehat{FGT}_i(\alpha, t) = N_i^{-1} \left[\sum_{j \in s_i} \widehat{FGT}_{ij}(\alpha, t) + \sum_{j \in r_i} \widehat{FGT}_{ij}(\alpha, t) \right],$$

where the unknown part for out-of-sample observations can be estimated using unit-level predictions $\hat{\mu}_{ij}$ and residuals e^*_{ij} by:

$$\sum_{j \in r_i} \widehat{FGT}_{ij}(\alpha, t) = n_i^{-1} \sum_{j \in s_i} \sum_{k \in r_i} \left(\frac{t - \hat{y}^*_{i(kj)}}{t} \right)^{\alpha} I(\hat{y}^*_{i(kj)} \leq t)$$

3.3.4 Uncertainty estimation

A discussion on the precision of area-level indicators necessitates reliable uncertainty estimates. An analytical assessment of uncertainty for area-level indicators is a challenging task even in the basic scenario of unit-level LMMs with block diagonal covariance matrices under unknown variance components (González-Manteiga et al., 2008; Rao and Molina, 2015). There exist partly-analytical approximations for means and totals (Prasad and Rao, 1990; Datta and Lahiri, 2000), however, for the determination of MSE estimates under complex model settings or for non-linear indicators, bootstrap schemes provide a suitable alternative (Hall and Maiti, 2006; González-Manteiga et al., 2008; Chambers and Chandra, 2013). We propose two flexible non-parametric bootstrap schemes (random effects block bootstrap (REB) and wild) for the estimation of domain-specific MSEs for economic and inequality indicators.

A major difference between the two bootstrap schemes roots in the generation of bootstrap populations. The REB bootstrap for non-linear indicators builds on the non-parametric bootstrap introduced by Chambers and Chandra (2013) modified for MERFs by Krennmair and Schmid (2022). The second bootstrap (wild) is inspired by Rojas-Perilla et al. (2020) and exclusively relies on centred OOB-residuals and a specific matching-scheme introduced by Feng et al. (2011) to build needed bootstrap populations.

The REB bootstrap captures the dependence-structure of the data and uncertainty introduced by the estimation of Model 3.1. Empirical residuals only depend on the correct specification of the mean behaviour function f of the model. The constructed bootstrap population requires an unbiased estimate on the residual-variance. However, the variance under the model, $\hat{\sigma}^2_\epsilon$ is positively biased, as it includes excess uncertainty regarding the initial estimation of the random forest $\hat{f}$ from the data (Mendez and Lohr, 2011). We maintain the extrapolation of this uncertainty, captured in the naive residuals, before the bootstrap population is simulated. Krennmair and Schmid (2022) scale and centre the empirical residuals by a bias-corrected residual variance (Mendez and Lohr, 2011) and eliminate uncertainty from the estimation of $\hat{f}$. Details regarding this bias-adjusted estimator for the residual variance of σ^2_ϵ can be found in Krennmair and Schmid (2022) and Section C.1.2 in the Appendix. Remaining steps of the proposed

bootstrap are as follows:

1. For given $\hat{f}()$, calculate the marginal residuals $z_{ij} = y_{ij} - \hat{f}(\mathbf{x}_{ij})$.

2. Using the marginal residuals $\hat{z}_{ij}$, compute level-2 residuals for each area by

$$\bar{z}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} z_{ij} \quad \text{for} \quad i = 1, ..., D$$

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

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

   (a) Sample independently with replacement from the scaled and centred level-1 and level-2 residuals:

$$z^{(b)}_{ij} = \text{srswr}(\hat{z}^c_{ij}, N) \quad \text{and} \quad \bar{z}^{(b)}_i = \text{srswr}(\bar{z}^c, D).$$

   (b) Calculate the bootstrap population as $y^{(b)}_{ij} = \hat{f}(\mathbf{x}_{ij}) + \bar{z}^{(b)}_i + z^{(b)}_{ij}$ and calculate the true bootstrap population indicator of interest $\delta^{(b)}_i$ for $i = 1, ..., D$.

   (c) For each bootstrap population (b), draw a bootstrap sample with the same n_i as the original sample. Use the bootstrap sample to obtain estimates $\hat{f}^{(b)}()$ and $\hat{v}^{(b)}_i$ as discussed in Section 3.3.1. Obtain estimates for indicators of interest $\hat{\delta}^{(b)}_i$ from estimated CDFs as discussed in Section 3.3.2.

5. Using the B bootstrap samples, the MSE estimator is obtained as follows:

$$\widehat{MSE}_i = B^{-1}\sum_{b=1}^{B}\left(\delta^{(b)}_i - \hat{\delta}^{(b)}_i\right)^2.$$

For the non-parametric wild bootstrap, we generate the bootstrap populations based on centred OOB-unit-level residuals and the empirical distribution of area-specific random effects. Details of the proposed bootstrap are as follows:

1. For given $\hat{f}()$, calculate OOB-residuals $e^{\text{OOB}}_{ij} = y_{ij} - \hat{f}(\mathbf{x}_{ij})^{\text{OOB}} - \hat{v}_i$ and save random effect elements $\hat{v}$.

2. Centre residuals e^{OOB}_{ij} and random effects $\hat{v}$ and denote them by $e^{\text{OOB}^*}_{ij}$ and $\hat{v}^*$.

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

   (a) Generate $v^{(b)}_i \overset{\text{iid}}{\sim} \text{srswr}(\hat{v}^*, D)$.

   (b) Calculate unit-level predictor values $\eta^{(b)}_{ij} = \hat{f}(\mathbf{x}_{ij}) + v^{(b)}_i$.

(c) Match $\eta_{ij}^{(b)}$ with the set of estimated unit-level predictors from the sample $\{\hat{\eta}_k | k \in s\}$ by finding the corresponding index $\tilde{k}$ solving $\min_{k \in s} |\eta_{ij}^{(b)} - \hat{\eta}_k|$.

(d) Generate weights w, where w is a simple two-point mass distribution with probabilities 0.5 at $w = 1$ and $w = -1$. This distribution satisfies the conditions in Feng et al. (2011).

(e) Generate the bootstrap population: $y_{ij}^{(b)} = \hat{f}(\mathbf{x}_{ij}) + \hat{v}_i^{(b)} + w_k |e_{\tilde{k}}^{\text{OOB}*(b)}|$ and compute bootstrap population indicators of interest $\delta_i^{(b)}$ for $i = 1, ..., D$.

(f) For each bootstrap population, draw a bootstrap sample with the same n_i as the original sample. Use the bootstrap sample to obtain estimates $\hat{f}^{(b)}()$ and $\hat{v}_i^{(b)}$ as discussed in Section 3.3.1. Obtain estimates for indicators of interest $\hat{\delta}_i^{(b)}$ from estimated CDFs as discussed in Section 3.3.2.

4. Using the B bootstrap samples, the MSE estimator for the indicator of interest is obtained as follows:

$$\widehat{MSE}_i = B^{-1} \sum_{b=1}^{B} \left(\delta_i^{(b)} - \hat{\delta}_i^{(b)} \right)^2 .$$

3.3.5 Distributional assumptions and transformation strategies

Our aim to introduce MERFs for the estimation of non-linear indicators follows the postulate of Efron (2020) to focus on the scientific applicability of machine learning methods for statistical subdisciplines. In the context of SAE, this leads inevitably to a discussion on trade-offs between model-flexibility and required control for dependency structures of survey data. In 'traditional' applications of SAE, domain-specific dependency is captured by random intercepts. Accordingly, we clarify consequences of the semi-parametric formulation of Model 3.1. The EM algorithm bridges concepts and exploits a Gaussian likelihood function to ensure the convergence towards a local maximum within the parameter space for required variance components σ_v and σ_ϵ (Hajjem et al., 2014). In Krennmair and Schmid (2022), we argue that the normality assumption on error terms ensures the existence of a closed-form solution of the integral over the Gaussian likelihood to calculate random effects, however, is not affecting the non-parametric estimation of fixed effects.

We aim to inform a transparent discussion and motivate further research between the spheres of traditional parametric SAE and the application of predictive algorithms by delivering empirical arguments along two dimensions: a) completely neglecting structural aspects of survey data, when using tree-based algorithms and b) exploring effects of transformation strategies on the dependent variable. The modification towards a fully non-parametric formulation of Model 3.1, for instance using discrete mixtures (Marino et al., 2019), is subject to further research. The idea of neglecting structural components follows the conjecture that increased predictive capabilities of random forests sufficiently capture patterns of area-level variations in the fixed effects part of the model. The idea of transformation strategies follows the general paradigm of fulfilling Gaussian assumptions on the transformed scale for the structural part of the model. For a), Predictor 3.4 reduces to $\hat{\mu}_{ij} = \hat{f}(\mathbf{x}_{ij})$, which makes the application of the EM algorithm obsolete. The estimation of the area-level CDF is straight-forward by Equation

3.6 and we refer to following poverty estimates based on this approach as RF. For b) we transform y_{ij} by $y_{ij}^{\mathrm{t}} = log(y_{ij})$ and obtain estimates for $\hat{f}^{\mathrm{t}}(), \hat{v}_i^{\mathrm{t}}, \hat{\sigma}_v^{2,\mathrm{t}}, \hat{\sigma}_\epsilon^{2,\mathrm{t}}$ as well as the estimated CDF $\hat{F}_i^{\mathrm{t},*}$ on the transformed scale. Indicators are calculated on the inverted scale of smeared values and we refer to this estimator as MERFlog.

A complete answer to the aspects raised in this section exceeds the purpose and capacity of this paper. Nevertheless, we will provide empirical arguments and discuss consequences of the two paradigms as part of the following design-based simulation.

3.4 Design-based simulation

Design-based simulations serve as realistic and controlled experiments to assess the performance of proposed methods for point and uncertainty estimates. We additionally test our method in a model-based simulation, which can be found in the Appendix. The following discussion of simulation results provides empirical evidence for the conceptual questions on distributional and structural assumptions raised in Section 3.3.5. Moreover, the analysis of performance highlights comparative efficiency advantages of our method, which helps to contextualize the estimates for our case study on the spatial distribution of poverty for the state of Veracruz in Section 3.5.

The variable *inglabpc* describes earned per capita income from work and exists in the census and in the survey data. Although *inglabpc* covers only one aspect of household income and deviates from the desired income definition, it is highly correlated to the target variable *ictpc*. Thus, our design-based simulation with the variable *inglabpc* is highly effective to assess the quality of our proposed estimates. We sample $M = 500$ independent samples from the fixed population of our census data set. Each pseudo-survey sample comprises the same number of in-sample households and shares identical domain-specific properties with the original survey data set as described by Table 3.1. True values of area-level HCR and PGAP are calculated based on census data for *inglabpc*.

We compare the performance of MERFs to established SAE methods that allow for the estimation of economic and inequality indicators. In particular, we juxtapose results to the EBP (Molina and Rao, 2010), the log-transformed EBP (EBPlog) and the EBP under data-driven Box-Cox transformation (EBPbc) by Rojas-Perilla et al. (2020). The EBP estimates serve as LMM-baseline for the estimation of non-linear indicators. The EBPlog mitigates deviations from Gaussian model assumptions and the EBPbc extends the transformation perspective by incorporating flexible and accurate transformations. Differences in the performance of estimates between the EBP, EBPlog and the EBPbc showcase advantages of the insurance to distributional violations of models by (data-driven) transformations. While differences between the MERF and the LMM-based methods highlight advantages of robustness due to less restrictive model assumptions and implicit model-selection. We additionally use RFlog and MERFlog to provide empirical evidence to the discussion in Section 3.3.5. RFlog trains a random forest on the log-transformed target variable and calculates indicators based on inverted values of domain-specific CDFs. Differences in the performance of tree-based approaches provide intuition on the importance of structural a priori model-specifications and the combined use of

transformation strategies in the the context of machine learning methods within the paradigm of SAE. Overall, we aim to show that our proposed estimator, which bridges concepts of flexibility and the incorporation of structural information, has comparative advantages over 'traditional' methods for estimation of area-level poverty indicators.

For the computational realization of the design-based simulation, we rely on the R-packages **emdi** (Kreutzmann et al., 2019) for the EBP, EBPlog and the EBPbc and package **SAEforest** (Krennmair, 2022) for estimates of our proposed MERF-approach and associated tree-based competitors. We monitor the convergence of the MERF-algorithm introduced in Section 3.3.2 with a precision of $1e^{-5}$ in relative difference of the GLL-criterion and keep the default of 500 trees for each forest. Based on three times repeated 5-fold cross-validation on the original survey data, we use 4 randomly drawn covariates as potential split-candidates at each tree's node (mtry) and a minimum of 5 observations in final nodes (min.node.size). The replication for the MSE bootstrap procedures are set to $B = 200$. For the EBP competitors, we do BIC-based stepwise selection on the (transformed) depend variable and keep the model fixed throughout the simulation. Details on the covariates of the optimally selected model are summarized in Table 3.1.

Metrics of evaluation for the competing methods are the empirical root mean squared error (RMSE) and the Bias for each specific indicator in area i ($\hat{\delta}_i^{\text{method}}$):

$$\begin{aligned} RMSE(\hat{\delta}_i^{\text{method}}) &= \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\delta}_i^{\text{method(m)}} - \delta_i^{(m)}\right)^2} \\ Bias(\hat{\delta}_i^{\text{method}}) &= \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\delta}_i^{\text{method(m)}} - \delta_i^{(m)}\right), \end{aligned} \tag{3.7}$$

where $\delta_i^{(m)}$ defines the true value of the indicator for area i in simulation round m.

We will start our discussion on the performance of point estimates focusing on the mean and median values of RMSE for HCR in Table 3.3. With respect to the total amount of 212 areas, our proposed MERF method has the lowest RMSE in mean and median terms. In general tree-based methods outperform the LMM-based competitors based on lower RMSE in mean and median terms. The data-driven EBPbc outperforms the EBPlog, however, interestingly the EBP without transformations delivers the best overall results in the class of LMM competitors. The results for the LMM-based methods are interesting and can partly be explained by the fact that transformations introduce bias. Usually, efficiency loss is outweighed by gains following the fulfilment of Gaussian assumptions. Comparing the HCR results to the PGAP for the EBP indicates that the untransformed EBP reproduces the distribution around the poverty line, however, completely fails to provide suitable information on other parts of the distribution needed to derive the PGAP.

We observe that the untransformed forests, outperform the log-transformed alternatives. This finding underlines initial claims on the resiliency to distributional assumptions on error terms. Moreover, smearing based on untransformed (i.e original scale) residuals leads to more accurate estimates for area-level CDFs compared to estimates based on scaled and back trans-

formed CDF estimates especially for higher quantiles. Our second major observation is that modelling the structural dependencies using random intercepts is rewarded by more accurate estimates in the transformed and untransformed case, although the comparative advantages are smaller in magnitude compared to unit-level mean and median squared prediction errors discussed in Table 3.2 from Section 3.2. The general observations on all areas are not affected by a detailed focus on in- and out-of-sample areas. Most interestingly, the RF and the RFlog have marginally lower RMSEs compared to the MERF and MERFlog for in-sample domains, which reverses for the majority of areas, i.e. the out-of-sample domains. The simulation results indicate that neglecting the structural knowledge on dependencies leads to overfitting on the implicit 'training' set , i.e. the survey data. Explicitly modelling the dependency structure of survey data leads to overall more reliable estimates.

Table 3.3: Mean and median for RMSE and Bias over total, in- and out-of-sample areas for point estimates of indicators HCR and PGAP.

		RMSE				BIAS			
		HCR		PGAP		HCR		PGAP	
		Median	Mean	Median	Mean	Median	Mean	Median	Mean
Total	EBP	0.0888	0.0916	0.4139	0.4239	0.0821	0.0656	0.4119	0.4223
	EBPbc	0.0975	0.1014	0.0519	0.0547	0.0950	0.0914	0.0496	0.0483
	EBPlog	0.1148	0.1166	0.0515	0.0537	0.1123	0.1092	0.0496	0.0462
	RFlog	0.0870	0.0912	0.0357	0.0408	0.0838	0.0759	0.0305	0.0214
	MERFlog	0.0870	0.0911	0.0351	0.0409	0.0785	0.0691	0.0258	0.0137
	RF	0.0808	0.0870	0.0303	0.0399	0.0726	0.0562	0.0152	0.0005
	MERF	0.0777	0.0854	0.0294	0.0420	0.0559	0.0364	0.0090	-0.0086
In-sample	EBP	0.0960	0.0999	0.3286	0.3270	0.0942	0.0913	0.3258	0.3247
	EBPbc	0.1004	0.1062	0.0482	0.0507	0.0974	0.1017	0.0459	0.0479
	EBPlog	0.1168	0.1220	0.0481	0.0504	0.1135	0.1179	0.0460	0.0473
	RFlog	0.0875	0.0921	0.0330	0.0356	0.0870	0.0875	0.0309	0.0300
	MERFlog	0.0970	0.0998	0.0338	0.0367	0.0889	0.0917	0.0267	0.0305
	RF	0.0775	0.0855	0.0241	0.0288	0.0761	0.0762	0.0200	0.0158
	MERF	0.0832	0.0884	0.0243	0.0300	0.0753	0.0755	0.0161	0.0163
Out-of-sample	EBP	0.0835	0.0884	0.4414	0.4604	0.0710	0.0559	0.4406	0.4591
	EBPbc	0.0953	0.0996	0.0526	0.0562	0.0937	0.0876	0.0511	0.0484
	EBPlog	0.1128	0.1145	0.0517	0.0549	0.1106	0.1060	0.0505	0.0458
	RFlog	0.0869	0.0909	0.0380	0.0427	0.0817	0.0715	0.0304	0.0181
	MERFlog	0.0828	0.0878	0.0356	0.0425	0.0755	0.0606	0.0232	0.0074
	RF	0.0827	0.0876	0.0328	0.0441	0.0680	0.0486	0.0134	-0.0053
	MERF	0.0740	0.0843	0.0320	0.0466	0.0410	0.0217	0.0024	-0.0180

Focusing on the more complex indicator of PGAP, we see a substantial improvement of transformation (and tree-based) strategies compared to estimates based on the untransformed EBP. Comparably to the results of the HCR, the tree-based competitors perform better than the LMM-alternatives resulting in lower levels of RMSEs. In comparison to the results of the HCR, we observe similar patterns between transformed and untransformed forests. Again, the RF has the lowest in-sample RMSE, however, for out-of-sample the MERF outperforms the RF in median terms. Comparing the performance of RF and MERF actuates the importance of modelling structural dependencies using random intercepts, although differences are marginal for the PGAP. Depending on the indicator, small differences in the RMSE between RF and MERF can also partly be explained by occasionally low intraclass correlations throughout the $M = 500$ simulation runs, as they vary approximately between a minimum of 0.005 and a

maximum of 0.11 with a median of around 0.04.

Apart from the RMSE, the Bias of area-level poverty indicators is a central aspect of quality. Table 3.3 reports that for the HCR, MERF estimates exhibit the lowest Bias in mean and median terms. Additionally, we observe a noticeable difference of about 30% reduction in Bias between the RF and the MERF over all areas and examine that the Bias of transformed approaches is relatively higher. For the PGAP, we observe similar patterns as compared to the HCR. Differences in Bias for transformed and untransformed LMM competitors are more pronounced compared to tree-based alternatives and confirm the necessity for suitable transformations for LMM-based approaches.

Overall, the comparison of point estimates indicates that MERFs perform competitively well in real-data applications and produce highly accurate results for the majority of areas. Their superior performance compared to LMM-based alternatives for non-linear indicators is in line with the observations for area-level means (Krennmair and Schmid, 2022). Although the tree-based methods without random intercept (RF and RFlog) perform competitively, we conclude that structure matters in terms of protection against in-sample overfitting and clearly in terms of reduced Bias for estimates. From the presented observations, there is no intention to use RFs or MERFs with a transformation for the estimation of HCR or PGAP for the example of Veracruz. Nevertheless, we observed that transformations in the context of MERFs and the estimation of area-level CDFs show comparative advantages for the estimation of lower quantiles. Poverty indicators, such as the PGAP and the HCR, however, mainly benefit from the preservation of the general shape of the distribution and a correct determination of the conditional mean, which is depended on a precise estimation of higher quantiles and extreme values. This is appropriately achieved by the proposed non-parametric generation of area-level CDFs with smearing. Nevertheless, the joint investigation of smearing under transformation and non-parametric procedures is subject to further research.

Table 3.4: Mean and median for relative RMSE and relative Bias of the estimated RMSE over total, in- and out-of-sample areas for point estimates of indicators HCR and PGAP.

		RRMSE_RMSE				RB_RMSE			
	[%]	HCR		PGAP		HCR		PGAP	
		Median	Mean	Median	Mean	Median	Mean	Median	Mean
Total	REB	40.42	89.86	45.84	103.85	9.84	65.33	14.65	76.96
	wild	42.20	98.37	42.66	109.05	5.79	69.59	5.21	77.09
In-sample	REB	26.84	36.19	29.75	43.80	-16.25	-1.91	4.99	22.14
	wild	36.73	41.96	34.03	45.08	-24.97	-11.11	-21.32	2.91
Out-of-sample	REB	44.04	110.08	53.49	126.46	35.60	90.65	21.01	97.61
	wild	45.24	119.61	50.83	133.14	37.21	99.99	23.86	105.03

We introduce a non-parametric REB and a wild bootstrap scheme in Section 3.3.4. The evaluation of MSE estimators based on real-world scenarios is computationally demanding and the most challenging and transparent way of assessing performance and reliability. Table 3.4 reports the RMSE and relative Bias of estimated RMSEs for the HCR and the PGAP. To determine the relative Bias and relative RMSE, we treat the empirical RMSE over Monte Carlo simulation rounds as basis of true values. Interestingly, the RRMSE_RMSE for the HCR and

the PGAP lie within the same range for total, in- and out-of-sample domains. The levels of RRMSE_RMSE are comparable for the wild as well as the REB bootstrap scheme. The large difference between mean and median values is attributed to heterogeneity of areas and outliers among the sampled and unsampled domains throughout the simulation rounds.

Regarding the RB_RMSE, we observe moderate levels of overestimation in median terms for the HCR and the PGAP for all domains. The wild bootstrap scheme reports moderate underestimation for in-sample areas for the HCR and the PGAP in median terms. The levels of RB_RMSEs for in-sample estimates align to comparable design-based studies within the field (Rojas-Perilla et al., 2020; Marchetti and Tzavidis, 2021), the evaluation of our proposed MSE procedures among the unsampled areas is challenging. The median levels of RB_RMSEs for HCR and PGAP for both MSE bootstrap schemes are acceptable and indicate overestimation, however, mean values appear to be extreme. Going into detail, we identify about 10% of out-of-sample areas that exhibit extreme variability throughout the simulation rounds. Interestingly, our proposed bootstrap estimators counter these issues with a tendency of overestimation, which is a more beneficial property than systematic underestimation in challenging data settings. We aim to provide a transparent discussion and realistic presentation of our method. Thus, we use most challenging scenarios to provide practitioners with a guidance on realistic properties of our methods. Complementing these results for the MSE estimators, we provide a model-based simulation in the Appendix to focus on the specific behaviour of our estimator in terms of controlled simulation experiments.

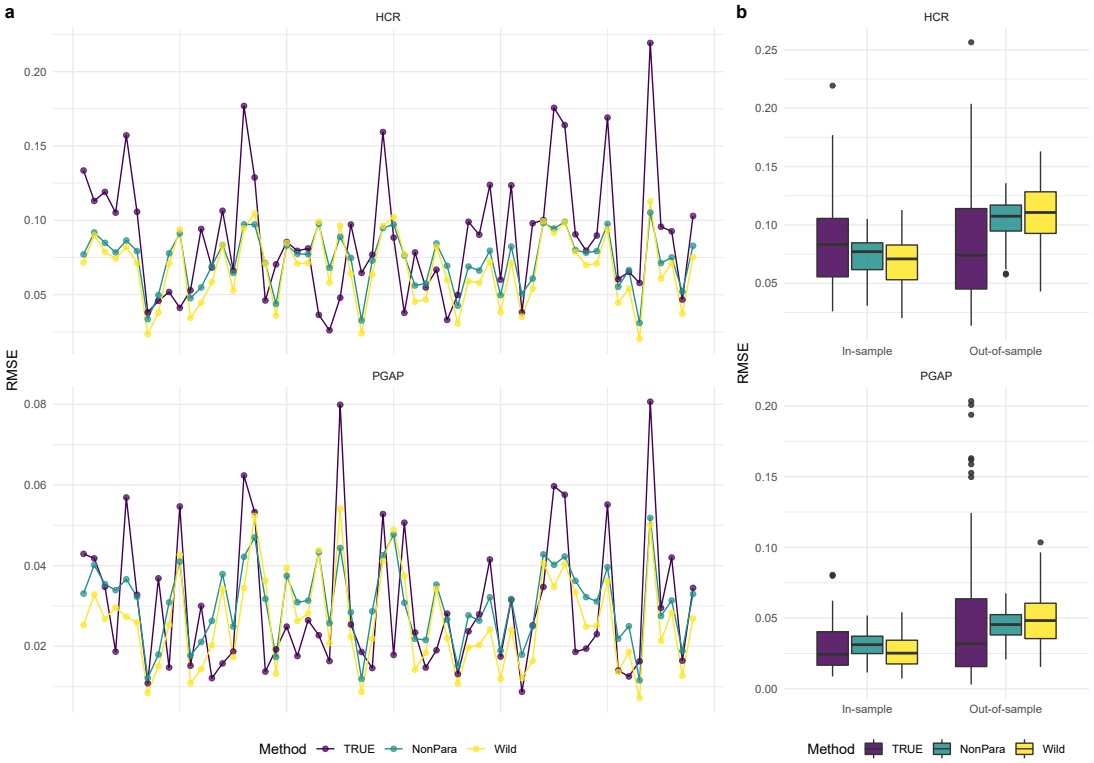


Figure 3.2: a) RMSE tracking properties for in-sample domains; b) summary of aggregated RMSEs over in- and out-of-sample domains.

Figure 3.2a reports the tracking properties of the wild and REB MSE procedure, visualizing

their estimated RMSE in comparison to the true RMSE for in-sample areas. We observe no systematic deviations and especially for the PGAP we report excellent tracking properties. Figure 3.2b summarizes the aggregated RMSEs over all in- and out-of-sample areas. Overall, we conclude that the wild and the REB bootstrap have a tendency to overestimate and thus deliver reliable, however, conservative MSE estimates in the context of realistic disaggregated poverty data applications. The wild and non-parametric procedure deliver comparable results.

3.5 Application and discussion of results

A major constraint for the spatial depiction of poverty in Veracruz is the lack of survey data for a majority of municipalities. In Section 3.2, we additionally maintain that the small sample sizes affect the precision and reliability of directs estimates. Model-based SAE improves the precision of in-sample estimates and provides empirical evidence for unsampled areas. Given our interest in subregional (non-linear) poverty indicators of HCR and PGAP based on the highly skewed target variable *ictpc*, we suggest the use of modelling techniques that handle non-normality and exhibit robustness against model-failure (Jiang and Rao, 2020). Particularly, we focus on the EBPbc and the proposed MERF as described in Section 3.3. We focus on two advanced modelling techniques, that stem from two alternative perspectives of predictive modelling in SAE. While the EBPbc remains within the paradigm of LMMs and uses data-driven parameter determination to find the optimal transformations ensuring Gaussian assumptions, MERFs bridge concepts of non-parametric flexibility and a priori model-specifications of dependency structures.

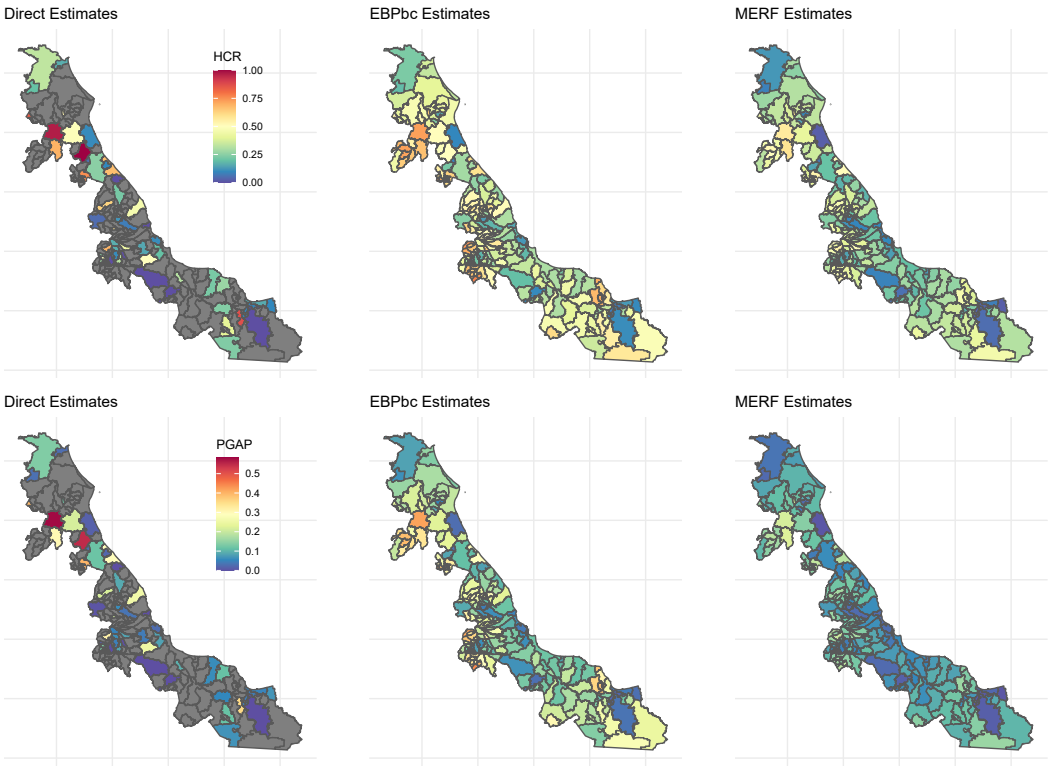


Figure 3.3: Estimated poverty based on household per capita income *ictpc* for the state of Veracruz based on direct estimates, EBPbc and MERF.

Figure 3.3 reports direct estimates, EBPbc and MERF estimates for the HCR and the PGAP. Referring to the direct estimates, we observe that sampled municipalities are located rather uniformly from north to the south and that the direct estimates are insufficient to identify spatial clusters of poverty. Inspecting the estimates based on the EBPbc and the MERF provides a complete poverty map. The juxtaposition of results reveals that MERF estimates for HCR and PGAP exhibit lower levels compared to the EBPbc. A comparison between direct estimates and the results from the two model-based variants confirms that high poverty areas remain unchanged in direct as well as model-based estimates, however, the EBPbc and MERF provide more balanced estimates. While the levels of HCR from EBPbc and MERF are comparable, PGAP estimates from MERFs are lower and show less variation. The results from the design-based simulation in Section 3.4 demonstrate more accurate estimates for municipality-level PGAP from MERFs for in- and out-of-sample areas. As the correlation between *inglabpc* and our target variable *ictpc* is high, we rely on estimates produced by the MERF. Following the empirical evidence, we infer that spatial patterns of poverty exist, however, the poverty intensity measured by the PGAP is moderate and relatively balanced throughout Veracruz.

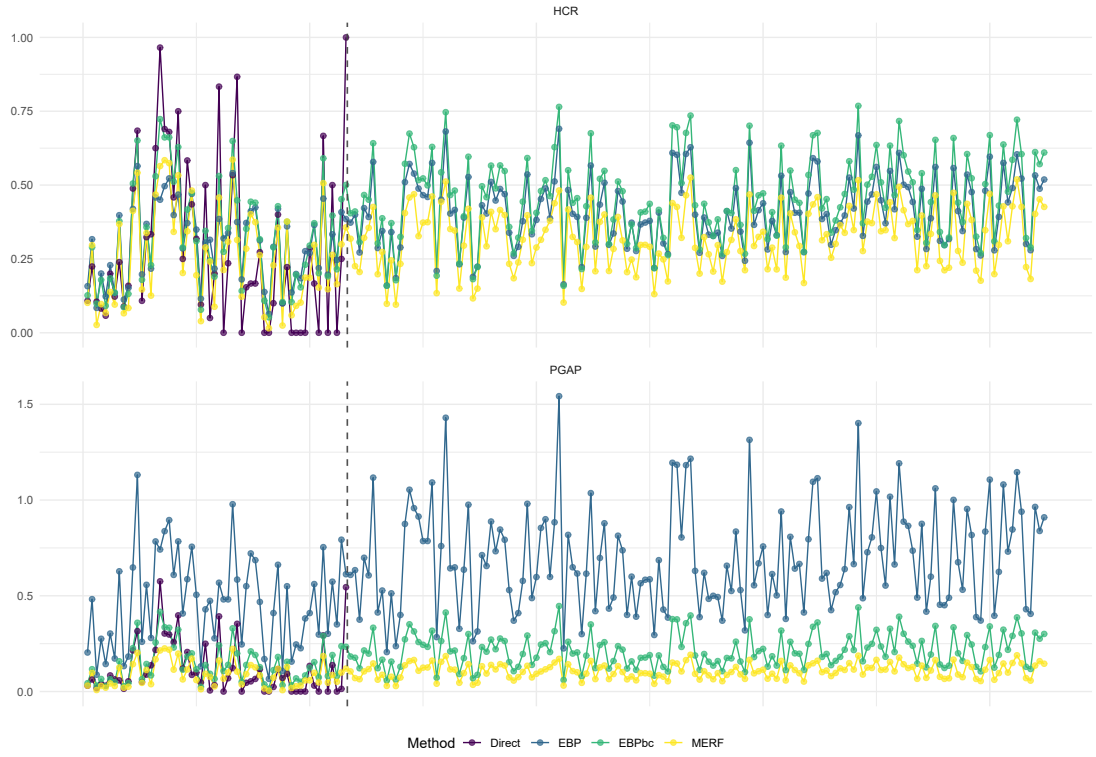


Figure 3.4: Estimated poverty based on household per capita income *ictpc* for the state of Veracruz based on direct estimates, the EBPbc and the MERF ordered by decreasing area sample sizes.

Figure 3.4 complements the discussion on the results of poverty estimates for the 212 municipalities. We order the 58 in-sample estimates by decreasing sample size to compare direct and model-based estimates in detail. Additionally, we add results from an untransformed EBP to highlight effects of data-driven transformations as well as the MERF's flexibility and robustness to model-failure. In the case of the HCR, we observe great similarities between model-

based and direct estimates for areas with higher sample sizes. The coherence between direct and model-based estimates reduces as expected with decreasing sample sizes. The EBP without transformation overestimates the PGAP drastically, which manifests the discussed inferior performance in the design-based simulation in Section 3.4.

The model-based poverty maps in Figure 3.3 provide empirical evidence for patterns of spatial distribution of poverty. Based on the HCR we identify several clusters of poverty. Going from north to south, we observe three major concentrations of poverty: firstly in the north-western regions sharing the border with the State of Hidalgo; secondly in the centre-western region Las Montañas sharing its border to Puebla and finally in the coastal region in the south-east, but northern to the harbour city of Coatzacoalcos. These three poverty clusters share the geographical characteristics of being in mountainous regions including the highest mountain Pico de Orizaba (5636m). The coastal regions and municipalities including the major cities, such as the capital Xalapa, are characterized by overall lower levels of poverty. Especially the coastal regions benefit from the economic opportunities of agriculture, tourism and petrochemical industry.

National statistical offices and multilateral organizations require empirical evidence of point estimates and associated information on the statistical reliability of indicators. The MSE of domain-specific indicators is used to construct important measures such as coverage rates, confidence intervals or CVs. The estimated MSE for each method is reported in Table 3.5 and was produced using the non-parametric REB bootstrap from Section 3.3.5 for the MERF. Variances of the direct estimates rely on the naive bootstrap (Alfons and Templ, 2013) and MSE estimations for EBPs are based on the parametric bootstrap introduced by Rojas-Perilla et al. (2020). All uncertainty measures were produced using $B = 200$ replications.

Table 3.5: MSE estimates for in- and out-of-sample domains of poverty indicators for competing methods.

MSE		HCR		PGAP	
		Median	Mean	Median	Mean
In-sample	Direct	0.0075	0.0104	0.0010	0.0024
	EBP	0.0031	0.0032	0.0156	0.0168
	EBPlog	0.0034	0.0036	0.0011	0.0011
	EBPbc	0.0032	0.0035	0.0011	0.0012
	MERF	0.0049	0.0052	0.0006	0.0008
Out-of-sample	EBP	0.0109	0.0107	0.0726	0.0806
	EBPlog	0.0145	0.0141	0.0059	0.0061
	EBPbc	0.0160	0.0155	0.0070	0.0073
	MERF	0.0244	0.0255	0.0037	0.0042

The MSEs of the HCR from Table 3.5 for in-sample domains report a substantial reduction in uncertainty for all model-based methods in mean and median terms compared to associated uncertainty of direct estimates. Interestingly, the MERFs exhibit higher levels of MSEs compared to the LMM-based alternatives for HCR. Given the superior precision of MERF point estimates in the design-based simulation in Section 3.4 and the fact that the RB_RMSE of the REB bootstrap signals moderate underestimation, we infer that the small MSE values of the EBP competitors may stem from underestimation. Focusing on the estimated uncertainty of PGAP indicators, we observe that the non-parametric REB bootstrap results report the low-

est values among all competitors for in- and out-of-sample domains. From the discussion of uncertainty estimates in the design-based simulation Section 3.4 reported in Table 3.4, we observe tendencies of moderate overestimation on the MSE values. Thus, we treat the reported MSEs as conservative upper bounds. A comparison between the variance of direct estimates and the MSEs of the poverty indicators indicate model-failure of the untransformed EBP for the estimation of domain-specific PGAPs in Veracruz. Also this observation is in line with the results from the design-based simulation.

3.6 Conclusion

In this paper, we propose MERFs for the estimation of disaggregated (non-linear) poverty indicators. In addition, we aim to inform a transparent methodological discussion on discrepancies between existing traditions of SAE and new emerging methods, such as (tree-based) machine learning methods and their contribution for poverty mapping. We maintain that our proposed estimators for point and uncertainty estimates meet modern requirements of SAE, including robustness against model-failure (Jiang and Rao, 2020). In a broader sense, this paper aims to introduce predictive methods to SAE by a critical scrutiny on statistical and practical requirements and with distinct focus on scientific applicability (Efron, 2020). Moreover, we aim to provide reliable empirical instruments to monitor the progress on disaggregated progress for the SDGs.

We introduce the case study on the Mexican state of Veracruz and motivate the necessity to use model-based SAE to provide empirical evidence for the spatial distribution of poverty. Following alternative perspectives to identify the best predictive model linking survey and census data, we focus on the semi-parametric unit-level model of MERFs and their subsequent extension to estimate area-level CDFs. We estimate poverty indicators of HCR and PGAP, which are predominantly used to measure progress on the eradication of poverty (SDG1). We complement approaches of point estimates by the introduction of two MSE bootstrap schemes: the non-parametric REB bootstrap and the wild bootstrap. We start a discussion on the necessity of bridging two concepts for the estimation of poverty indicators and subsequently evaluate our proposed methods for point and MSE estimates in a design-based simulation. The design-based simulation manifests that our proposed estimates have comparative advantages in terms of RMSE and Bias compared to 'traditional' model-based approaches. Our approach is characterized by robustness against distributional violations of normality and shows advantages in the presence of unknown and potentially complex interactions of covariates. We use a design-based simulation to evaluate and compare our proposed methods and additionally provide results of a model-based simulation in the Appendix.

Further research from a methodological perspective are extensions using MERFs for non-linear indicators to capture multidimensional aspects of poverty. Our general discussion and our methodological framework for the production of (non-linear) poverty indicators additionally aims to motivate the use of other predictive machine learning approaches such as Boosting, Support Vector Machines or Bayesian additive regression trees. Depending on the flexibility of the predictors, fully non-parametric formulations capturing area-level dependency structures

impose an interesting direction for further research. From the perspective of advanced applications, questions concerning the estimation of disaggregated indicators in the absence of census data using geospatial information become increasingly relevant (Wardrop et al., 2018). Accordingly, the use of alternative big data covariates (Marchetti et al., 2015; Schmid et al., 2017) is an interesting application for our proposed method and the needed monitoring of progress on SDGs on subregional levels. This research direction potentially extends the methodological discussion on cultures of SAE (e.g. traditional vs. flexible predictive models) to dimensions of 'traditional' (e.g. census and administrative) and alternative data sources (e.g. telephone or geospatial data) and corresponding best practices.

Acknowledgements

The authors are grateful to CONEVAL for providing the data used in empirical work. The views set out in this paper are those of the authors and do not reflect the official opinion of CONEVAL. The numerical results are not official estimates and are only produced for illustrating the methods. Additionally, the authors would like to thank the HPC Service of ZEDAT, Freie Universität Berlin, for computing time.

Appendix C

C.1 Technical appendix

C.1.1 Algorithm of Monte Carlo approximation to the CDF

Let δ_i be a (poverty) indicator of interest for area i and $h()$ is a function that calculates this indicator. We can write $\delta_i = h(y_{s_i} \cup y_{r_i})$, where s_i are sampled and r_i are non-sampled observations. We summarize unknown parameters for our assumed super-population model in Section 3.3.1 in parameter c. We can write a predictor for δ_i by:

$$\hat{\delta}_i = h(y_{s_i} \cup E(y_{r_i}|y_s; \hat{c}))$$

The following Monte Carlo simulation can be used to approximate the unknown indicator of interest by simulating the estimated conditional distribution of y_{ij} for units outside the sample:

1. For given $\hat{f}()$ calculate the model residuals $\hat{e}_{ij} = y_{ij} - \hat{f}(\mathbf{x}_{ij}) - \hat{v}_i$ and save random effects $\hat{v}$ from each area i and variance components $\hat{\sigma}_\epsilon$ and $\hat{\sigma}_v$.

2. Scale $\hat{e}_{ij}$ by $\hat{\sigma}_\epsilon$ and centre the residuals. Scale the vector of random effects $\hat{v}$ by $\hat{\sigma}_v$. Denote adjusted variance components as $\hat{e}^c_{ij}$ and $\hat{v}^c$ respectively. Additionally calculate $\gamma_i = \dfrac{\hat{\sigma}_v}{\hat{\sigma}_v + \hat{\sigma}_\epsilon / n_i}$.

3. For $m = 1, ..., M$:

   (a) Sample independently with replacement from the empirical distributions of $\hat{e}^c_{ij}$ and $\hat{v}^c$:

   $$e^{(m)}_{ij} = \text{srswr}(\hat{e}^c_{ij}, N) \quad \text{and} \quad \tilde{v}^{(m)}_i = \text{srswr}(\hat{v}^c(1 - \gamma_i), D).$$

   (b) Simulate the bootstrap population as $y^{(m)}_{ij} = \hat{f}(\mathbf{x}_{ij}) + \hat{v}_i + \tilde{v}^{(m)}_i + e^{(m)}_{ij}$.

   (c) Determine the indicators of interest $\delta^{(b)}_i$ for $i = 1, ..., D$.

4. Using the M simulation rounds, final estimates of indicators are:

$$\hat{\delta}_i = M^{-1} \sum_{m=1}^{M} \hat{\delta}^{(m)}_i$$

Please note that for out-of-sample observations, $\hat{v} = 0$ and the shrinkage factor γ_i becomes 1 such that the marginal distribution is simulated. Additionally, in many applications of SAE the sampling fraction is very small and it is impossible to clearly separate covariate information in $\mathbf{x}_{ij}$ between sampled and non sampled population units. Thus, we take the whole available covariate information such that our estimate reduces to $\hat{\delta}_i = h(E(y_{ik}|y_s; \hat{c}))$.

C.1.2 Bias-adjustment of residual variance

The estimation of variance components in Step 2 (d) of the MERF-algorithm in Section 3.3.1 for $\hat{\sigma}^2_\epsilon$ and $\hat{\sigma}^2_v$ is obtained by taking the expectation of maximum likelihood estimators given the data. Although $\hat{\sigma}^2_\epsilon$ is a naive estimator within the discussed framework, it cannot be considered as an unbiased estimator for the variance σ^2_ϵ of the unit-level errors ϵ_{ij}. Breiman (2001a) maintains that the sum of squared residuals from OOB-predictions is a valid estimator for the squared prediction error of new individual observations. However, as an estimator of the residual variance under the model, $\hat{\sigma}^2_\epsilon$ is positively biased, as it includes uncertainty regarding the estimation of the random forest $\hat{f}$. Following Mendez and Lohr (2011) we use a bias-adjusted estimator for the residual variance σ^2_ϵ from Model 3.1 using a bootstrap bias-correction. The essential steps to obtain the corrected residual variance are summarized as follows:

1. Use the OOB-predictions $\hat{f}(\mathbf{x}_{ij})^{\text{OOB}}$ from the final model $\hat{f}()$ after convergence of the algorithm.

2. Generate B bootstrap samples $y^\star_{ij,(b)} = \hat{f}(\mathbf{x}_{ij})^{\text{OOB}} + \epsilon^\star_{ij,(b)}$, where the values $\epsilon^\star_{ij,(b)}$ are sampled with replacement from the centred marginal residuals $\hat{e}_{ij} = y_{ij} - \hat{f}(\mathbf{x}_{ij})^{\text{OOB}}$.

3. Recompute $\hat{f}(\mathbf{x}_{ij})^{\text{OOB}}_{(b)}$ using a random forest with $y^\star_{ij,(b)}$ as dependent variable.

4. Estimate the correction-term $K(\hat{f})$ by:

$$\hat{K}(\hat{f}) = B^{-1} \sum_{b=1}^{B} \left[\hat{f}(\mathbf{x}_{ij})^{\text{OOB}} - \hat{f}(\mathbf{x}_{ij})^{\text{OOB}}_{(b)} \right]^2.$$

The bias-corrected estimator for the residual variance is then given by:

$$\hat{\sigma}^2_{bc,\epsilon} = \hat{\sigma}^2_\epsilon - \hat{K}(\hat{f}). \tag{C.1}$$

Table C.1: Explanation of variables ranked by importance of the random forest including all variables from the optimal model of the EBPbc after model-selection from 39 potential covariates.

Variable name	Explanation	EBPbc-model
ictpc	Total household income per capita.	
escol_rel_hog	Average relative amount of schooling standardized by age and sex of household members.	✓
bienes	Availability of goods in the household.	✓
actcom	Assets in the household.	✓
jnived	Formal education of the household's head.	
jaesc	Average years of schooling of household members.	✓
jtocup	Occupation type.	✓
pcocup	Percentage of employed household members.	
est_calidad_vivienda	Classification of dwellings based on their conditions and services.	
jexp	Years of working experience of household's head.	
jedad	Household member's age.	
tdep	Household members under 16 years and over 65, divided by the members between 16 to 64.	
pcpering	Percentage of income earners in the household.	✓
tam_loc	Number of inhabitants of closest town.	✓
jsector	Indicator for sector of activity of head and/or spouse.	✓
iingmuj	Identifies households with female prevalence income.	✓
muj_hog	Total number of women in the household.	✓
bengob	Identifies whether the household receives government income compensation.	✓
tam_hog	Number of household members.	✓
autoconsumo	Household with the presence of a member working in the primary sector.	✓
nalfab	Total number of literates in the household.	✓
totocup_hog	Total number of employed in the household.	✓
muj16_notrb_hog	Women over 16 years of age not working at home.	✓
rururb	Indicates whether household location is rural or urban.	✓
hijas_hog	Total number of daughters in the household.	✓
jubi	Presence of retired people or pensioners in the household.	✓
clase_hog	Identifies the type of household.	✓
tmor_hog	Measures the mortality rate in the household.	✓
alimc_2	Identifies the households in which members report days without food.	✓
pob_ind	Identifies whether household members speak indigenous languages.	✓

C.2 Model-based simulation

The model-based simulation compares the performance of area-level estimates for the HCR and PGAP. We use the same competing methods from Section 3.4 (EBP, EBPlog, EBPbc, MERF, MERFlog, RF and RFlog). Overall, we assess the quality of point and uncertainty estimates from our proposed methodology and highlight advantages of robustness in controlled scenarios of model-misspecification.

The simulation-setting follows survey-sample properties of our case study for Veracruz. We assume a finite population P of size $N = 58000$ with $D = 58$ separate areas $P_1, ..., P_D$ of equal size $N_i = 1000$, which corresponds to the median area size of original census data from Table C.1. We utilize the 58 small in-sample areas to generate stratified random samples, which resemble our empirical data. The total sample size is $n = \sum_{i=1}^{D} n_i = 1453$, which coincides with the design-based simulation (Section 3.4) and the application (Section 3.5). The poverty line for all scenarios is set to 30% of y, which corresponds to the mean of poverty over all 212 domains from Section 3.5.

Table C.2: Model-based simulation scenarios

Scenario	Model	x_1	x_2	μ	v	ϵ
Normal	$y = 10000 - 500x_1 - 500x_2 + v + \epsilon$	$N(\mu, 3^2)$	$N(\mu, 3^2)$	$unif(-1, 1)$	$N(0, 500^2)$	$N(0, 1000^2)$
Interaction-lowICC	$y = 5000 + 500x_1x_2 + 250x_2^2 + v + \epsilon$	$N(\mu, 1^2)$	$N(\mu, 2^2)$	$unif(-1, 1)$	$N(0, 200^2)$	$N(0, 1000^2)$
Interaction-highICC	$y = 5000 + 500x_1x_2 + 250x_2^2 + v + \epsilon$	$N(\mu, 1^2)$	$N(\mu, 2^2)$	$unif(-1, 1)$	$N(0, 500^2)$	$N(0, 1000^2)$
Logscale	$y = exp(7.5 - 0.25x_1 - 0.5x_2 + v + \epsilon)$	$N(\mu, 1^2)$	$N(\mu, 1^2)$	$unif(1, 2)$	$N(0, 0.1^2)$	$N(0, 0.25^2)$
Normal-GB2	$y = 10000 - 1000x_1 - 500x_2 + v + 0.5\epsilon$	$N(\mu, 3^2)$	$N(\mu, 3^2)$	$unif(0, 1)$	$N(0, 1250^2)$	$GB2(2.2, 2500, 18, 1.46)$

The simulation comprises five scenarios denoted as *Normal*, *Interaction-lowICC*, *Interaction-highICC*, *Logscale* and *Normal-GB2*. We repeat each scenario independently $M = 500$ times. Two major dimensions of model-failure contextualize the following discussion on competing estimates for various indicators: firstly, the presence of skewed data delineated by non-normal error-terms or log-transformed data and secondly, the presence of unknown non-linear interactions between covariates. The baseline Scenario *Normal* meets model assumptions for LMMs. The Scenarios *Interaction-lowICC* and *Interaction-highICC* share a similar error-structure with Scenario *Normal*, however, the fixed effects include quadratic terms and interactions, showcasing comparative advantages of automated model-selection of MERFs. The two interaction scenarios emphasize the necessity to model structural dependencies using random intercepts and differ in estimated intraclass correlations (ICCs) from MERFs of approximately 2% for Scenario *Interaction-lowICC* and around 15% for Scenario *Interaction-highICC*. In Section 3.3.5 we discuss transformation approaches for MERFs. Scenario *Logscale* demonstrates benefits of transformations and Scenario *Normal-GB2* combines the linear additive structure of LMMs with GB2-distributed unit-level errors. Both scenarios aim to promote the use of transformation strategies to meet distributional assumptions of LMMs. Additionally, we highlight robustness properties of our proposed methods for point and uncertainty estimates. Further details on the data-generating process for each scenario are provided in Table C.2.

C.2.1 Discussion of point estimates

Comparably to Section 3.4, we analyse the performance of competing methods based on RMSE and Bias of HCR and PGAP. In the baseline Scenario *Normal*, we observe that the EBP shows the lowest RMSE for HCR and PGAP and results are similar to the EBPbc. This demonstrates benefits of the flexible data-driven Box-Cox transformation in comparison to the fixed transformation of EBPlog. Secondly, we observe that the MERF outperforms tree-based competitors and that levels of RMSE are competitively close to the LMM-based alternatives, which mirror the data-generating process. Scenarios *Interaction-lowICC* and *Interaction-highICC* demonstrate comparative advantages of flexibility and automated model-selection of the RF and MERF. The MERF reports the smallest RMSE for all indicators. Irrespectively of the degree of ICC, we observe efficiency gains of modelling structural dependencies using random intercepts, which become more serious with increasing ICC. We observe the lowest RMSE among the LMM competitors for the EBPbc. Although the EBPbc fails to replicate the correct model, it outperforms the RF for HCR and PGAP in both Scenarios (*Interaction-lowICC*, *Interaction-highICC*), which is attributed to the importance of accounting for domain-specific dependencies.

Table C.3: Mean and median for RMSE and Bias in model-based scenarios for point estimates of indicators HCR and PGAP.

		RMSE				BIAS			
		HCR		PGAP		HCR		PGAP	
		Median	Mean	Median	Mean	Median	Mean	Median	Mean
Normal	EBP	0.0341	0.0356	0.0080	0.0084	-0.0026	-0.0026	-0.0005	-0.0005
	EBPbc	0.0341	0.0356	0.0081	0.0084	-0.0027	-0.0027	-0.0005	-0.0005
	EBPlog	0.0636	0.0642	0.0133	0.0142	0.0487	0.0491	0.0027	0.0027
	RFlog	0.0745	0.0729	0.0177	0.0173	0.0164	0.0161	0.0016	0.0014
	MERFlog	0.0415	0.0424	0.0093	0.0102	0.0144	0.0144	0.0005	0.0004
	RF	0.0729	0.0711	0.0177	0.0173	0.0017	0.0011	0.0007	0.0005
	MERF	0.0355	0.0370	0.0085	0.0088	-0.0012	-0.0014	-0.0007	-0.0009
Interaction-lowICC	EBP	0.0509	0.0508	0.0376	0.0372	-0.0075	-0.0061	0.0347	0.0345
	EBPbc	0.0513	0.0518	0.0129	0.0129	0.0166	0.0176	0.0052	0.0052
	EBPlog	0.0528	0.0528	0.0177	0.0178	0.0125	0.0135	0.0079	0.0079
	RFlog	0.0534	0.0532	0.0126	0.0125	0.0134	0.0135	0.0009	0.0007
	MERFlog	0.0464	0.0453	0.0106	0.0109	0.0131	0.0132	0.0001	-0.0000
	RF	0.0518	0.0513	0.0133	0.0132	-0.0012	-0.0013	0.0045	0.0043
	MERF	0.0425	0.0427	0.0110	0.0109	-0.0020	-0.0027	0.0036	0.0031
Interaction-highICC	EBP	0.0898	0.0905	0.0418	0.0404	-0.0080	-0.0071	0.0328	0.0324
	EBPbc	0.0810	0.0823	0.0223	0.0222	0.0179	0.0181	0.0052	0.0053
	EBPlog	0.0838	0.0850	0.0265	0.0266	0.0187	0.0189	0.0082	0.0079
	RFlog	0.1188	0.1173	0.0309	0.0310	0.0164	0.0163	0.0015	0.0014
	MERFlog	0.0642	0.0660	0.0171	0.0184	0.0188	0.0183	-0.0002	-0.0005
	RF	0.1181	0.1162	0.0316	0.0313	0.0002	-0.0001	0.0054	0.0051
	MERF	0.0592	0.0617	0.0159	0.0166	-0.0025	-0.0037	0.0018	0.0016
Logscale	EBP	0.0506	0.0509	0.1429	0.1432	-0.0204	-0.0212	0.1269	0.1277
	EBPbc	0.0340	0.0347	0.0136	0.0141	0.0005	0.0006	0.0002	0.0004
	EBPlog	0.0340	0.0348	0.0136	0.0141	0.0006	0.0007	0.0002	0.0004
	RFlog	0.0567	0.0565	0.0231	0.0231	0.0041	0.0039	0.0019	0.0019
	MERFlog	0.0356	0.0361	0.0145	0.0147	0.0027	0.0024	0.0007	0.0006
	RF	0.0579	0.0582	0.0257	0.0257	-0.0070	-0.0072	0.0112	0.0112
	MERF	0.0430	0.0443	0.0218	0.0218	-0.0089	-0.0091	0.0114	0.0113
Normal-GB2	EBP	0.0621	0.0647	0.0485	0.0493	0.0023	0.0017	0.0311	0.0315
	EBPbc	0.0698	0.0712	0.0296	0.0307	0.0455	0.0452	0.0141	0.0138
	EBPlog	0.2260	0.2263	0.2423	0.2418	0.2084	0.2083	0.2367	0.2366
	RFlog	0.1144	0.1140	0.0651	0.0649	0.0415	0.0413	0.0353	0.0353
	MERFlog	0.1178	0.1189	0.0742	0.0739	0.0588	0.0589	0.0462	0.0457
	RF	0.1074	0.1067	0.0542	0.0542	0.0100	0.0101	-0.0035	-0.0037
	MERF	0.0616	0.0638	0.0326	0.0340	0.0051	0.0048	-0.0064	-0.0062

Scenario *Logscale* highlights benefits of (log-)transformation approaches. As expected the EBPlog estimates HCR and PGAP most efficiently, closely followed by the EBPbc. The EBPlog has the lowest RMSE among all competitors for HCR and PGAP. Comparing the EBP and EBPlog for the PGAP reveals that the magnitude of efficiency gains from correct transformation is larger compared to tree-based alternatives (e.g comparing MERFlog to the MERF). The final Scenario *Normal-GB2* addresses skewed data due to GB2-distributed error terms. Comparing the RMSE of EBP and EBPlog for HCR and PGAP demonstrates a case, where the wrong transformation is less efficient than neglecting Gaussian assumptions. Benefits of EBPbc become visible for the complex PGAP indicator. Interestingly, MERFlog and RFlog outperform the EBPlog, which again highlights the general adaptability of (ME)RFs. The MERF and the EBPbc perform competitively well with lowest RMSE of MERFs for HCR and the EBPbc for the PGAP.

The Bias of estimates is an essential detail of the RMSE. We do not observe severe discrepancies in the performance of competitors between RMSE and their Bias. MERFS are charac-

terized by relatively low levels of Bias throughout all scenarios. Except for Scenario *Logscale*, EBPbc results are associated with more (positive as well as negative) Bias compared to the MERF. This is particularly interesting as unit-level predictions from random forests usually achieve efficiency gains with introduced Bias. However, in our example this behaviour from unit-level predictions does not transfer to the Bias of constructed non-linear poverty domain indicators. Overall, the model-based results complement insights from the design-based simulation in Section 3.4 and demonstrate that MERFs serve as insurance to model-misspecification and are an agnostic competitor to 'traditional' SAE approaches balancing required flexibility and structural dependencies.

C.2.2 Discussion of MSE estimates

We use our five scenarios to assess the quality of proposed uncertainty estimators from Section 3.3.4. We determine relative Bias and relative RMSE by defining the empirical RMSE over Monte Carlo simulation rounds as basis for true values. Table C.4 reports mean and median percent of relative RMSE and Bias over all 58 areas and $M = 500$ simulation rounds. Starting with the RRMSE_RMSE, we observe that all values for the wild bootstrap exceed the non-parametric REB bootstrap for all indicators and scenarios. Regarding the RB_RMSE, we see that under Scenario *Normal* the REB bootstrap is essentially unbiased for HCR and PGAP and the wild bootstrap exhibits moderate overestimation. For the Scenarios *Interaction-lowICC* and *Interaction-highICC*, we observe moderate underestimation of the REB bootstrap for HCR and PGAP, which reduces with higher ICC. The wild bootstrap is a conservative uncertainty estimator for the PGAP in both scenarios. Scenario *Logscale* is constructed to showcase benefits of transformations. The REB bootstrap handles the challenging data scenario more reliably compared to the wild bootstrap exhibiting overestimation for both non-linear poverty indicators. Scenario *Normal-GB2* is characterized by a skewed distribution and extreme outliers. Both bootstrap schemes indicate negative RB_RMSE. Nevertheless, the non-parametric REB bootstrap appears to reproduce extreme distributional characteristics of error-terms better than the wild bootstrap, leading to less biased results, especially for the HCR.

Table C.4: MERF: Mean and median for relative RMSE and relative Bias of the estimated RMSE for HCR and PGAP in model-based scenarios.

		RRMSE_RMSE				RB_RMSE			
	[%]	HCR		PGAP		HCR		PGAP	
		Median	Mean	Median	Mean	Median	Mean	Median	Mean
Normal	REB	7.02	7.31	8.16	8.36	-0.09	0.08	0.10	-0.14
	wild	8.43	8.88	10.25	10.40	3.59	4.02	4.98	5.10
Interaction-lowICC	REB	16.68	16.28	14.32	14.13	-3.46	-3.65	-0.92	-0.80
	wild	18.78	19.00	22.63	22.88	-11.56	-11.67	18.25	18.02
Interaction-highICC	REB	6.93	7.60	9.82	10.41	-1.41	-1.92	-0.19	-0.72
	wild	10.28	10.80	15.88	16.52	-7.73	-7.81	12.32	11.90
Logscale	REB	28.82	29.13	36.51	37.90	16.97	16.00	17.88	15.93
	wild	40.05	40.41	57.08	57.74	31.38	30.82	42.47	41.40
Normal-GB2	REB	12.26	12.60	28.26	28.21	-8.62	-8.43	-27.02	-26.83
	wild	18.88	19.16	36.01	36.25	-17.35	-17.46	-35.24	-35.33

Chapter 4

The R package SAEforest

4.1 Introduction

Reliably measurable metrics are imperative to monitor demographic, economic and social development. Typically national statistical offices produce and administer elaborate statistical indicators based on survey data. With increasing availability of (alternative) data sources, research institutes and multilateral organizations aim to quantify precise information at a finer geographical resolution. The terms 'domain' or 'area' define separate entities within a joint population, such as (but not limited to) districts within a country. Many surveys are designed to produce accurate estimates at national (or sub-national levels). With deliberated disaggregation of domains, the accuracy of direct estimates decreases with domain-specific sample sizes and model-based small area estimation (SAE) offers promising tools. By combining auxiliary data sources via models with survey data, SAE methods implicitly increase the effective precision of domain-specific indicators of a target variable. Overviews of existing methods for SAE are found in Pfeffermann (2013), Rao and Molina (2015) or Tzavidis et al. (2018).

Predominant models for SAE are conceptualized within the regression-setting and the majority relies on linear mixed models (LMM) to account for the hierarchical structure of survey data (Rao and Molina, 2015). The predictive performance of parametric models relies on the fulfilment of (Gaussian) model assumptions, but economic and inequality data is often highly skewed and characterized by deviations from the normal distribution. Jiang and Rao (2020) maintain that methodological improvements in SAE must focus on robustification of models against model-failure (e.g. providing insurances against model-misspecification, valid variable selection and the effective handling of outliers). Optimality results of parametric LMMs depend on the validity of model assumptions, which becomes challenging for applications dealing with social and economic inequality data. Existing strategies to cope with deviations from (Gaussian) assumptions are, for instance, (data-driven) transformation strategies of the dependent variable (Molina and Martín, 2018; Sugasawa and Kubokawa, 2019; Rojas-Perilla et al., 2020) or less restrictive assumptions on unit-level models (Diallo and Rao, 2018; Graf et al., 2019). In the presence of outliers, means can be determined using robustified LMMs (Sinha and Rao, 2009) or M-quantile approaches (Chambers and Tzavidis, 2006), which estimate non-linear indicators without a formal specification of random effects (Tzavidis et al., 2010; Marchetti and Tzavidis, 2021). Opsomer et al. (2008) use penalized splines regression

for the estimation of are-level means, dealing with non-linearities by treating spline coefficients as additional random effects.

Machine learning methods offer non-linear and nonparametric alternatives, combining excellent predictive performance and a reduced risk of model-misspecification. Krennmair and Schmid (2022) introduce mixed effects random forests (MERF) as versatile tools for applications in model-based SAE. MERFs combine advantages of regression forests (e.g. implicit model-selection and robust predictive performance in the presence of outliers) with the ability to model hierarchical dependencies. Package **SAEforest** provides a coherent user-friendly framework facilitating the use of MERFs for the estimation of spatially disaggregated (non-) linear indicators and their respective uncertainty, measured by reliable mean squared errors (MSE).

In recent years, ongoing methodological contributions in (model-based) SAE are increasingly complemented by the development of open-source R-packages. I aim to give a comprehensive overview of existing SAE related packages on the Comprehensive R Archive Network (CRAN) focussing on unit-level models. Moreover, I aim to discuss existing packages dealing with random forests under dependent data sources, to motivate the functionality of the **SAEforest** package:

The package **sae** (Molina and Marhuenda, 2015) offers a suitable collection of SAE methods for point and uncertainty estimates for area and unit-level models. Package **emdi** (Kreutzmann et al., 2019) focusses on the estimation of disaggregated economic and inequality indicators (and respective uncertainty) and insures against model-misspecification implementing an EBP under data-driven transformations (Rojas-Perilla et al., 2020). The package treats the EBP by Molina and Rao (2010) as a special case and combines computationally efficient methods with a genuine workflow on data processing and presentation of results. Additional packages for unit-level survey data are package **JoSAE** (Breidenbach, 2018), which focuses on models coping with heteroskedasticity. From a Bayesian perspective, the package **hbsae** (Boonstra, 2022) combines functions for various unit- and area-level models, bridging frequentist and Bayesian perspectives. A complete Bayesian workflow for the estimation demographic and health indicators is found in package **SUMMER** (Li et al., 2021). Outlier-robust estimators from a Bayesian perspective are provided by package **robustsae** (Ghosh et al., 2016) and from a more frequentist perspective by **saeRobust** (Warnholz, 2018) or the **rsae** package Schoch (2014).

Existing packages for dependent data and tree-based machine learning methods are not concerned with topics of SAE and hardly focus on inference. The package **LongituRF** (Capitaine, 2020) bundles functions that allow for time-invariant covariance structures and rely on a semi-parametric unit-level mixed model for regression trees and forests. Although the primary focus of package **MixRF** (Wang and Chen, 2016) is the imputation of clustered and incomplete data, the package comprises a genuine function, with which MERFs can be estimated. Functions from package **RandomForestGLS** (Saha et al., 2021) model spatial random effects as Gaussian processes by developing dependency adjusted split-criteria handling dependent error processes similarly to generalized least squares. Package **splinetree** (Neufeld and Heggeseth, 2019) builds regression trees and random forests for longitudinal or dependent data using a

spline projection method.

The major aim of package **SAEforest** is the provision of a complete and coherent use of MERFs for SAE. Current packages with a focus on random forests for dependent data are not intended to estimate SAE indicators and associated measures of uncertainty. On the other hand, existing unit-level SAE packages neglect tree-based methods. The use of MERFs in SAE promotes general flexibility for domain-level predictions and package **SAEforest** combines methods on the estimation of point and MSE estimates for various indicators.

Implemented estimators rely on the empirical and methodological contributions introducing MERFs for SAE of means by Krennmair and Schmid (2022), for non-linear indicators by Krennmair et al. (2022a) as well as in the case of aggregated auxiliary information by Krennmair et al. (2022b). The flexibility of the package does not only stem from methodological aspects, but from the provision of a genuine workflow for practitioners of SAE. **SAEforest** puts emphasis on the integration of methods and generic functions that facilitate the summary and visualization of results. Additionally, predefined tools for diagnostics and the tuning of MERF hyper-parameters are available, such as the number of trees (`num.trees`) or the number of randomized split-candidates at each node (`mtry`). Implemented functions for MERFs are easily adaptable and allow for potential extensions to advanced patterns of correlation and multilevel structures.

The paper is organized as follows: Section 4.2 provides an overview of the statistical methodology used in the package. This includes a formal introduction to MERFs, details on the estimation of domain-level means with unit-level and aggregated covariates, as well as the estimation of non-linear indicators and corresponding MSEs. Section 4.3 describes data sources used as examples in the package. The core functionality of the package and its features are explained in Section 4.4. Section 4.5 summarizes methods and results and raises ideas for further research.

4.2 Statistical methodology

This section introduces a general mixed model enabling a simultaneous discussion of traditional LMM-based models in SAE, such as the nested error regression model of Battese et al. (1988) and semi-parametric interpretations, such as the model of Krennmair and Schmid (2022) using MERFs. Machine learning methods are popular alternatives for predictive modelling in various scientific disciplines (Varian, 2014; Efron, 2020). Tree-based data-driven prediction algorithms (such as random forests (Breiman, 2001b)) combine flexible modelling properties without explicit model assumption. Moreover, they identify complex higher-order relations in covariates and show robustness properties in the presences of outliers (Hastie et al., 2009; Biau and Scornet, 2016). Thus, random forests contribute to the robustification of models against model-failure (Jiang and Rao, 2020). In order to become a genuine tool for SAE, predictive data-driven procedures must meet basic premises of survey and inference theory, such as the handling of hierarchically dependent data structures and measures of uncertainty for produced indicators.

In the following sections, we will discuss the estimation of reliable domain-specific statis-

tical indicators from survey data using MERFs and focus on their respective MSEs. Additional emphasis lies on the estimation of area-level means without population micro-data. The methods introduced are illustrated as part of an example on synthetic Austrian income data in Section 4.4 and rely on the theoretical and empirical methods provided by Krennmair and Schmid (2022) and Krennmair et al. (2022b) for means and Krennmair et al. (2022a) for non-linear indicators.

4.2.1 A general mixed effects model for SAE and MERFs

We assume a finite population U of size N consisting of D domains $U_1, U_2, ..., U_D$ with $N_1, N_2, ..., N_D$ units, where index $i = 1, ..., D$ denotes respective areas. For every individual observation j in area i in the sample, we observe the continuous target variable y_{ij}. We draw sample s of size n from population U and sampled observations are assigned to D respective areas resulting in sample sizes $n_1, n_2, ..., n_D$. A sub-sample from area i is denoted by s_i and corresponding non-sampled observations are denoted by r_i. The p predictive covariates $\mathbf{x}_{ij} = (x_1, x_2, ..., x_p)^\intercal$ are assumed to be available for every unit within the sample s. The following general mixed effects regression model describes the relationship between $\mathbf{x}_{ij}$ and y_{ij}:

$$y_{ij} = f(\mathbf{x}_{ij}) + u_i + e_{ij} \quad \text{with} \quad u_i \sim N(0, \sigma_u^2) \quad \text{and} \quad e_{ij} \sim N(0, \sigma_e^2). \tag{4.1}$$

Function $f(\mathbf{x}_{ij})$ models the conditional mean of y_{ij} given $\mathbf{x}_{ij}$. The hierarchical structure of observations is captured by area-specific random intercepts u_i and we assume independence between u_i and unit-level errors e_{ij}.

For instance, defining $f(\mathbf{x}_{ij}) = \mathbf{x}_{ij}^\intercal \beta$ with $\beta = (\beta_1, ..., \beta_p)^\intercal$ resembles the definition of the nested error regression model by Battese et al. (1988), which serves as basis for a majority of unit-level SAE-models. Well known examples are the EBP by Molina and Rao (2010) or the EBP under data-driven transformations by Rojas-Perilla et al. (2020). Under known optimality results of LMMs, optimal estimates of fixed effects $\hat{\beta}$ and variance components $\hat{\sigma}_u^2, \hat{\sigma}_e^2$ are obtained by maximum likelihood (ML) or restricted maximum likelihood (REML) (Rao and Molina, 2015).

We combine predictive advantages of random forests with the ability to model hierarchical structures of survey data with random effects by defining f in Model 4.1 to be a random forest (Breiman, 2001a). Resulting MERFs rely on a procedure reminiscent of the EM-algorithm (Hajjem et al., 2014) to obtain optimal estimates on model components $\hat{f}$, $\hat{u}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$. The proposed MERF algorithm fits parameters for Model 4.1 (where f is a random forest) by iteratively estimating a) the forest function, assuming the random effects term to be correct and b) the random effects part, assuming the Out-of-Bag-predictions (OOB-predictions) from the forest to be correct. OOB-predictions correspond to the unused observations in the internal bootstrap step prior to the construction of each forest's sub-tree (Breiman, 2001a; Biau and Scornet, 2016). We estimate variance components $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_u^2$ by implicitly taking the expectation of ML estimators given the data. Computationally, the MERF algorithm is implemented in the function `MERFranger` of **SAEforest**. Note that step a) is realized using package **ranger** (Wright and Ziegler, 2017), while the estimation of variance components and random effects builds on package **lme4** (Bates et al., 2015). The convergence of the algorithm is monitored

by marginal changes of log-likelihood of the composite semi-parametric model. For further methodological details, we refer to Krennmair and Schmid (2022). The proposed estimator for model-based predictions is given by:

$$\hat{\mu}_{ij}^{\text{MERF}} = \hat{f}(\mathbf{x}_{ij}) + \hat{u}_i = \hat{f}(\mathbf{x}_{ij}) + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i} \left(\frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - \hat{f}^{\text{OOB}}(\mathbf{x}_{ij})) \right). \tag{4.2}$$

4.2.2 Flexible domain prediction of means under unit-level and aggregated covariates

The predictions $\hat{\mu}_{ij}^{\text{MERF}}$ (4.2) depend on auxiliary unit-level information to estimate unit-level conditional means for the continuous dependent variable. In the context of SAE, however, researchers are mainly interested in estimating and mapping indicators such as area-level means or metrics measuring income deprivation and inequality (Rao and Molina, 2015). For now, we will focus on the construction of area-level means depending on the availability of unit-level or aggregated auxiliary covariate information. The construction of domain-specific cumulative distribution functions (CDFs) from which non-linear indicators can be obtained will be discussed in Section 4.2.3.

For unit-level (i.e. $\mathbf{x}_{ij}$) supplementary data (usually census or administrative data), we calculate the mean-estimator for each area i by:

$$\hat{\mu}_i^{\text{MERF}} = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \hat{u}_i = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i} \left(\frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - \hat{f}^{\text{OOB}}(\mathbf{x}_{ij})) \right), \tag{4.3}$$

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

We exploit the fact that random forest estimates of the fixed part $\hat{f}()$ express the conditional mean at unit-level and that $\hat{u}_i$ is the best linear unbiased predictor (BLUP) for the linear part of Model 4.1 (Krennmair and Schmid, 2022). For non-sampled areas, the proposed estimator for the area-level mean reduces to the fixed part from the random forest:

$$\hat{\mu}_i = \bar{\hat{f}}(\mathbf{x}_{ij}).$$

The access to auxiliary population micro-data for covariates imposes a limitation for researchers and practitioners. As a direct consequence of non-linearity and non-continuity of random forests, we observe that $f(\bar{\mathbf{x}}_i) \neq \bar{f}_i(\mathbf{x}_{ij})$ and aggregated auxiliary information cannot directly be processed into predictions on μ_i in Equation 4.2. Krennmair et al. (2022b) solve this issue by incorporating aggregate population-level covariate information through calibration weights w_{ij}, balancing unit-level predictions from MERFs in Equation 4.2 in coherence with the area-wise covariate means from census data. In short, the estimator for area-level means under limited auxiliary information is given by:

$$\hat{\mu}_i^{\text{MERFagg}} = \sum_{j=1}^{n_i} \hat{w}_{ij} \left[\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i \right]. \tag{4.4}$$

The optimal estimates from survey data for required model components $\hat{f}$ and $\hat{u}_i$ using the MERF algorithm are similar to Equation 4.2. The $\mathbf{x}_{ij}$ for Estimator 4.4 are unit-level covariates from the survey and population-level auxiliary information is incorporated through optimal calibration weights $\hat{w}_{ij}$ maximizing the profile empirical likelihood (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;
- $w_{ij} \geq 0$, preventing the cancellation of weights;
- $\sum_{j=1}^{n_i} w_{ij} = 1$, ensuring the normalization of weights.

The Lagrange multiplier method is suitable to find optimal weights (Owen, 1990, 2001) and Krennmair et al. (2022b) discuss technical conditions for the feasibility of solutions in the context of SAE and propose a best practice strategy that is implemented in this package.

Irrespectively of the quality of auxiliary data sources (aggregated or unit-level), the function `SAEforest_model` provides methods to assess the uncertainty of point estimates with domain-specific MSEs. The quantification of uncertainty of domain-indicators is challenging, yet essential for the assessment of reliability of area-level estimates. Approximating the analytical MSE of domain-level indicators with estimated variance components remains challenging even in the base scenario of LMMs with block diagonal covariance matrices (Prasad and Rao, 1990; Datta and Lahiri, 2000; González-Manteiga et al., 2008; Rao and Molina, 2015). Elaborate bootstrap-schemes for the estimation of MSEs are an established alternative (Hall and Maiti, 2006; González-Manteiga et al., 2008; Chambers and Chandra, 2013) and the preferred choice under our general mixed model.

We propose a nonparametric random effect block (REB) bootstrap for estimating the MSE of area-level means of sampled and unsampled domains. The major aim is the correct reproduction of dependence-structures of data and an incorporation of uncertainty introduced through the estimation of the MERF. The nonparametric generation and resampling of random components was originally introduced by Chambers and Chandra (2013). Krennmair and Schmid (2022) postulate the importance to resample centred and scaled empirical error components by a bias-adjusted residual variance introduced by Mendez and Lohr (2011) before constructing a bootstrap population. In short, the estimator of the residual variance under the MERF from Equation 4.2, ($\hat{\sigma}_\epsilon^2$) is positively biased as it includes excess uncertainty concerning the estimation of function $\hat{f}$. Further methodological and performance details are found in Krennmair and Schmid (2022). For cases of existing unit-level auxiliary covariates, we imitate the sampling process by random draws from the simulated bootstrap populations. In the presence of aggregated population-level data, we generate (pseudo-) true values by resampling error components only. This idea follows methodological principles of the bootstrap for finite populations introduced by González-Manteiga et al. (2008). For details, model-based simulations and examples, please see Krennmair et al. (2022b).

4.2.3 Non-linear indicators

The analysis of distributional aspects of consumption and income (in-) equality based on statistical indicators builds on a long tradition in statistical research (Atkinson, 1987; Cowell, 2011). In contrast to the estimation of domain-specific means, the model-based estimation of quantiles and (non-linear) poverty indicators requires information on the area-specific CDF of y_{ij}. Chambers and Dunstan (1986) (CD) combine a model for a finite-population CDF of y_{ij} with a smearing-argument (Duan, 1983) to develop a model-consistent estimator for a finite-population CDF from survey sample data. Tzavidis et al. (2010) introduce the CD-method within a general unit-level framework for SAE with a focus on the estimation of SAE means and quantiles in the context of a bias-adjusted alternative to the EBLUP and outlier-robust M-quantile estimators. Extensions towards poverty (Marchetti et al., 2012) and inequality indicators (Marchetti and Tzavidis, 2021) were investigated.

Rooted within the general unit-level framework of Tzavidis et al. (2010), Krennmair et al. (2022a) propose an estimator $F_i^*(t)$ for the area-specific CDF of y_{ij} using MERFs. Essentially, we extend the smearing method to $\hat{\mu}_{ij}$ as given by Estimator 4.2 using OOB-residuals $e_{ij}^* = y_{ij} - \hat{\mu}_{ij}^{\text{OOB}}$, where $\hat{\mu}_{ij}^{\text{OOB}} = \hat{f}^{\text{OOB}}(\mathbf{x}_{ij}) + \hat{u}_i$. OOB-residuals are a genuine choice for achieving more robust estimates of the CDF of MERFs, ensuring that these model-residuals e_{ij}^* mirror the estimated variance properties under Model 4.1. The estimator for $F_i^*(t)$ is given by:

$$\hat{F}_i^*(t) = N_i^{-1} \left[\sum_{j \in s_i} \mathbf{I}(y_{ij} \leq t) + n_i^{-1} \sum_{j \in s_i} \sum_{k \in r_i} \mathbf{I} \left(\hat{\mu}_{ik} + \underbrace{(y_{ij} - \hat{\mu}_{ij}^{\text{OOB}})}_{e_{ij}^*} \leq t \right) \right] \tag{4.5}$$

Smearing is computationally intensive and a Monte Carlo (MC) approximation to the area-specific CDF of y_{ij} provides an alternative. The MC-based approach draws conceptual parallels to the EBP (Molina and Rao, 2010), however, lacks theoretical foundation (Marchetti et al., 2012). Nevertheless, the MC approximation to Equation 4.5 is time-efficient and given a sufficiently high number of iterations (e.g. `B_MC = 200`) no obviously identifiable difference between point estimates for various indicators are observable. **SAEforest** provides both methods and recommends the use of the theoretically supported smearing approach as default.

Estimates for indicators δ_i are calculated from $\hat{F}_i^*(t)$ using a known function $h()$. Default indicators and corresponding functions $h()$ are defined in Table 4.1. Package **SAEforest** includes the $(10\%, 25\%, 50\%, 75\%, 90\%)$ quantiles as default indicators characterizing the distribution of y_{ij}. We additionally include common economic measures of poverty such as the head count ratio (Hcr) and the poverty gap (Pgap) (Foster et al., 1984) and inequality measures such as the Gini coefficient (Gini, 1912) and the Quintile share ratio (Qsr) (Eurostat, 2004). The Hcr defines the rate of being at risk of poverty, while the Pgap ratios the mean income shortfall of the poor to its respective poverty line. Both poverty indicators require a poverty threshold (z), which can be defined in absolute terms (e.g. numerical values of national poverty lines) or relative terms (e.g. defining a function depending on y_{ij}). Package **SAEforest** allows for both options. Focussing on distributional aspects, the Gini is a common measure summarizing inequality between 0 (absolute) equality and 1 (absolute inequality). While the Gini bundles

information on the whole distribution, the Qsr focusses on the relation between joint income (or consumption) of the 80 and 20 percent quantile. Additionally, users can use a custom function for arbitrary statistical indicators relying on input `Y` and threshold `z`. The example in Section 4.4.1 will discuss customizable features in detail.

Table 4.1: List of predefined population indicators in **SAEforest**. F_i is the empirical distribution function in domain i.

Indicator	Definition $h()$	Range
Mean$_i$	$\frac{\sum_{j=1}^{N_i} y_{ij}}{N_i}$	$\mathbb{R}$
Q$_{i,q}$	$F_i^{-1}(q) = \inf\{y_{ij} \in \mathbb{R} : F_i(y_{ij}) \geq q\}$	$\mathbb{R}$
Hcr$_i$	$\frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{I}(y_{ij} \leq z)$	$[0,1]$
Pgap$_i$	$\frac{1}{N_i} \sum_{j=1}^{N_i} (\frac{z - y_{ij}}{z}) \mathbf{I}(y_{ij} \leq z)$	$[0,1]$
Gini$_i$	$\frac{2 \sum_{j=1}^{N_i} j y_{ij}}{N_i \sum_{j=1}^{N_i} y_{ij}} - \frac{N_i + 1}{N_i}$	$[0,1]$
Qsr$_i$	$\frac{\sum_{j=1}^{N_i} \mathbf{I}(y_{ij} > Q_{i,0.8}) y_{ij}}{\sum_{j=1}^{N_i} \mathbf{I}(y_{ij} \leq Q_{i,0.2}) y_{ij}}$	$[0,1]$
custom$_i$	$g(y_{ij}, z)$	$\mathbb{R}$

Following the work of Krennmair et al. (2022a), the package provides two bootstrap schemes (`nonparametric` and `wild`), each applicable for the smearing and the MC-based versions. The major difference between the two bootstrap schemes is the generation of the bootstrap population. The nonparametric bootstrap prepares and resamples random components for its bootstrap population in the same way as described in Section 4.2.2 and subsequently calculates (non-linear) indicators from the simulated data. The wild bootstrap (`wild`) exclusively relies on centred OOB-residuals and a specific matching scheme between sampled and synthetic observations building the bootstrap population. Details and performance specifics for both procedures are found in Krennmair et al. (2022a).

4.3 Data set description

Typical applications of SAE comprise survey sample data on target variable y_{ij} and predictive variables $\mathbf{x}_{ij}$. Since existing auxiliary data sources (census or administrative/register data) do not include information on the target variable, auxiliary data sources strengthen estimates on disaggregated metrics of y_{ij} through a predictive model. As discussed in Section 4.2, we provide models, which handle auxiliary covariates of domain-specific individual observations or domain-level aggregates (e.g means). The exemplary datasets in this package include both types of information for illustrative purpose.

In general, this package uses data examples provided by package **emdi** (Kreutzmann et al., 2019). In short, the datasets comprise simulated synthetic data from the European Union Statistics on Income and Living Conditions (EU-SILC) for Austria from 2006. Although no conclusions regarding the official levels of inequality and poverty in Austrian districts must be in-

ferred, the simulated population micro-dataset `eusilcA_pop` exhibits realistic distributional characteristics. Originally, the `eusilcA_pop` data is a modification of the `eusilcP` data used in package **simFrame** (Alfons and Templ, 2013), which reports micro-data on the nine states as lowest geographical level. Kreutzmann et al. (2019) use publicly available sources, such as population sizes or income rankings of districts, to assign households to one of the 94 districts. Further details on the process of data synthetization can be found in Kreutzmann et al. (2019).

Focussing on social and economic inequality indicators, the target variable is the equivalized household income (`eqIncome`). For the construction of `eqIncome`, total household disposable income is divided by the equivalized household size (Hagenaars et al., 1994). Apart from domain-level identifiers for states (`state`) and the districts (`districts`), auxiliary variables are socio-demographic characteristics, such as gender or the receipt of state benefits. An overview of model covariates is provided in Table D.1 in the Appendix. The dataset `eusilcA_popAGG` comprises aggregated district-level means and is used for the illustration of Method 4.4 in Section 4.2.2. For the production of uncertainty estimates, Method 4.4 requires information on population-level domain sizes. Synthetic population sizes for Austrian districts are provided by `popnsize`.

The unit-level sample `eusilcA_smp` is drawn by stratified random sampling from the population dataset, where districts are defined as stratas. The resulting dataset comprises 1945 observations with domain-specific sample sizes ranging from 14 ("Lienz") to 200 for the Austrian capital ("Wien"). About 25 percent of domains are not covered by the survey dataset, additionally motivating the use of model-based SAE approaches. For the illustration of the mapping function `map_indicators`, we use a shape file for the Austrian districts of class `SpatialPolygonDataFrame` (Bivand et al., 2013), obtainable from package **emdi** (Kreutzmann et al., 2019).

4.4 Core functionality: the package

The statistical methods for point and MSE estimates from Section 4.2 are implemented in the main function `SAEforest_model`. The functionality of the package mirrors the proposed methodological flexibility of tree-based machine learning methods: firstly, depending on the available auxiliary data sources (aggregated or unit-level covariates) and the indicators of interest (means or non-linear indicators), domain-specific estimates are produced using `SAEforest_model`. Users must specify corresponding scenarios with options `meanOnly = TRUE` and/or `aggData = TRUE`. Resulting model objects can be checked by summary statistics and visual model diagnostics using the generic functions `summary` and `plot`. Function `tune_parameters` assesses potential improvements of the model by tuning model hyper-parameters. Finally, function `summarize_indicators` extracts final domain-specific estimates and function `map_indicators` visualizes and maps indicators upon request. Detailed examples on the functionality of proposed methods follow in the subsections below.

Generic functions of the package rely on S3 objects of class `SAEforest` (Chambers and Hastie, 1992). The function `SAEforest_model` wraps the basis function `MERFranger`.

Table 4.2: Details on inputs for main function `SAEforest_model`.

Input	Description	meanOnly = TRUE	meanOnly = FALSE
`Y`	Continuous target variable.	✓	✓
`X`	Matrix or data.frame of predictive covariates.	✓	✓
`dName`	Character of domain identifier.	✓	✓
`smp_data`	data.frame of survey sample data.	✓	✓
`pop_data`	data.frame of population-level covariates `X`.	✓	✓
`MSE`	Specification of uncertainty estimates. Currently available options are: `none`, `nonparametric` and for `meanOnly = F` additionally `wild`.	✓	✓
`importance`	Variable importance processed by **ranger**. Must be one of the following: "impurity", "impurity_corrected" or "permutation".	✓	✓
`initialRandomEffects`	Initial estimate of random effects. Defaults to 0.	✓	✓
`ErrorTolerance`	Value monitoring MERF algorithm's convergence. Defaults to 1e-04.	✓	✓
`MaxIterations`	Value specifying maximal amount of iterations for MERF algorithm. Defaults to 25.	✓	✓
`B`	Bootstrap replications for MSE estimation. Defaults to 100.	✓	✓
`B_adj`	Bootstrap replications for adjustment of residual variance. Defaults to 100.	✓	✓
`na.rm`	Logical. Whether missing values should be removed.	✓	✓
`...`	Additional parameters passed to `ranger`. Most important parameters are `mtry` (number of variables to possibly split at in each node), or `num.trees` (number of trees).	✓	✓
`aggData`	Logical. Whether aggregated covariate information is used.	✓	
`popnsize`	Information of population size of domains. Only needed if `aggData = TRUE` and MSE is requested.	✓	
`OOsample_obs`	Out-of-sample observations taken from the closest area. Only needed if `aggData = TRUE` with default set to 25.	✓	
`ADDsamp_obs`	Out-of-sample observations taken from the closest area if first iteration for the calculation of calibration weights fails. Only needed if `aggData = TRUE` with default set to 0.	✓	
`w_min`	Minimal number of covariates from which informative weights are calculated. Only needed if `aggData = TRUE`. Defaults to 3.	✓	
`threshold`	Set a custom threshold for indicators. The threshold can be a known numeric value or function of `Y`. Defaults to `NULL` resulting in 60% of median of `Y`.		✓
`custom_indicator`	A list of additional functions containing the indicators to be calculated. These functions must only depend on the target variable `Y` and the `threshold`. Defaults to `NULL`.		✓
`smearing`	Logical input indicating whether a smearing based approach or a MC-based version for point estimates is obtained. Defaults to `TRUE`.		✓
`B_MC`	Bootstrap populations to be generated for the MC version. Defaults to 100.		✓

The implementation of the MERF algorithm is done by a composite model of a random forest fitted by the package **ranger** (Wright and Ziegler, 2017) and random intercepts and corresponding variance components obtained by the package **lme4** (Bates et al., 2015). Thus, users benefit from the full functionality of both package environments including generic functions of respective classes `ranger` and `merMod`. Moreover, users can directly pass hyper-parameters

to the function `ranger` or choose alternative splitrules for trees. Although the basis function `MERFranger` is only addressed through wrapper functions for the average package user, we additionally provide the function to enable unit-level predictions under more advanced correlation and dependency structures. By this, we aim to facilitate further research and development using MERFs for SAE. For details, see `help(MERFranger)` or the methodology discussed in Krennmair and Schmid (2022).

4.4.1 Estimation of domain-level indicators

The following examples use the synthetic Austrian EU-SILC data discussed in Section 4.3. Firstly, we focus on the most ideal case including unit-level survey sample data and access to unit-level covariate data from a census to estimate the area-level mean. The information on the equivalized income is only measured in the survey data, but covariates `X_covar` are measured on survey and census level.

```
R> #Loading data
data("eusilcA_pop")
data("eusilcA_smp")

income <- eusilcA_smp$eqIncome
X_covar <- eusilcA_smp[,-c(1,16,17,18)]
```

This data scenario corresponds to Method 4.2. As we are only interested in the area-level mean, we specify option `meanOnly = TRUE` and define target variable `Y` and corresponding covariates in the sample `X = X_covar`. Input values for covariates `X` must be predictors only and we remove columns containing area-level codes and the target variable for the assignment `X = X_covar`. We explicitly denote `dName` to indicate separate areas for random intercepts and assign the survey dataset `smp_data` and the dataset comprising population-level information `pop_data`. For the current example, point estimates are sufficient and we specify `MSE = "none"`. As discussed in Section 4.2, the current implementation has an option to produce uncertainty estimates of area-level means with option `nonparametric` referring to the MSE procedures discussed in Krennmair and Schmid (2022). Dealing with unit-level population data, we keep the default of `aggData = FALSE`. Note that this option must be replaced by `TRUE` in the case of limited covariate information.

```
R> MERFmodel1 <- SAEforest_model(Y = income, X = X_covar,
+         dName = "district", smp_data = eusilcA_smp, pop_data =
+         eusilcA_pop, MSE = "none", meanOnly = TRUE,
+         aggData = FALSE)
```

Before we discuss model components and respective results, we focus on inputs for estimating more complex area-level indicators, such as quantiles or inequality indicators. Function `SAEforest_model` with option `meanOnly = FALSE` corresponds to the methodology explained in Section 4.2.3 and allows for further scenario-dependent inputs. The option `smearing` determines whether we want to construct a full smearing CDF or choose

a Monte-Carlo simulated marginal distribution of y_{ij}. Depending on computational feasibility, we advice the general use of smearing-based estimates due to its theoretical corroboration compared to the MC version. For MSE estimates, we have options `none, wild` or `nonparametric` as described in Krennmair et al. (2022a). The default indicators returned by `SAEforest_model` with option `meanOnly = FALSE` include the mean, median, quantiles (10%, 25%, 75% and 90%), Hcr, Pgap, Gini, and the Qsr. Users specify a custom threshold by passing a known numeric value or a function of `Y`. If the threshold is `NULL`, 60 % of the median of `Y` is taken as threshold. Additionally, `SAEforest_model` allows for custom indicators. In the following example, we constructed a new indicator, defining area-level maximum incomes. The input for `custom_indicator` must be a list of functions depending only on inputs `Y` and `threshold`.

```
R> MERFmodel2 <- SAEforest_model(Y = income, X = X_covar,
+         dName = "district", smp_data = eusilcA_smp,
+         pop_data = eusilcA_pop, smearing = FALSE,
+         meanOnly = FALSE, MSE = "nonparametric", B = 100,
+         mtry=5, num.trees = 500, threshold =
+         function(Y){0.5 * median(Y)}, custom_indicator =
+         list(my_max = function(Y, threshold){max(Y)}))
```

Function `SAEforest_model` allows to pass arguments directly to the function `ranger` using the generic three-dotted option (`...`). Most important inputs to specify a random forest are the number of randomized variables for each node split decision (`mtry`) or the overall number of trees (`num.trees`). Any option available for `ranger` (such as alternative split criteria) can be directly passed to the function. For details, see Wright and Ziegler (2017) and our discussion on tuning parameters in Section 4.4.3. Table 4.2 in the Appendix summarizes and explains the inputs for `SAEforest_model`.

Function `SAEforest_model` produces an output object of class `SAEforest`, which always includes at least four elements: (i) point estimates of specified regionally disaggregated indicators; (ii) a `MERFmodel` object including information on the model fit for fixed effects and random effects; (iii) MSE estimates if requested and `NULL` otherwise; (iv) the value of the adjusted standard deviation used in the MSE bootstrap or `NULL` otherwise. In the case of domain-level means under aggregated covariate information (`aggData = TRUE`), the object additionally includes an element, capturing the number of variables used in the weighting process from aggregated covariate information. Table 4.3 summarizes and explains individual components of `SAEforest` objects. Several generic methods are applicable and we firstly focus on model diagnostics produced by `summary` and `plot` in the following section.

4.4.2 Summary function and diagnostic plots

Function `summary` is an important generic method to obtain essential information on a fitted model object. An exemplary output from `summary` of a fitted model object of class `SAEforest` is given below:

```
R> summary(MERFmodel1)
```

```
Call:
SAEforest_model(Y = income, X = X_covar, dName = "district",
smp_data = eusilcA_smp, pop_data = eusilcA_pop, MSE = "none",
aggData = FALSE, importance = "impurity")

Domains
In-sample Out-of-sample Total
 70            24         94
Totals:
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

Random forest component:
Type:                             Regression
Number of trees:                         500
Number of independent variables:          14
Mtry:                                      3
Minimal node size:                         5
Variable importance mode:           impurity
Splitrule:                          variance
Rsquared (OOB):                      0.62036

Structural component of random effects:
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: Target ~ -1 + (1 | district)
Data: data
Offset: forest_preds
AIC      BIC      logLik  deviance  df.resid
39225.2  39236.3 -19610.6  39221.2     1943
Scaled residuals:
Min      1Q      Median    3Q     Max
-3.1425 -0.5243 -0.0577  0.4433 11.6832

Random effects:
Groups   Name        Variance Std.Dev.
district (Intercept) 12132734 3483
Residual             30771664 5547
Number of obs: 1945, groups:  district, 70
ICC:  0.2827853

Convergence of MERF algorithm:
Convergence achieved after 8 iterations.
A maximum of 25 iterations used and tolerance set to: 1e-04
Monitored Log-Likelihood:
-19546.21 -19572.14 -19588.23 -19592.72 -19604.67 -19599.86 -19609.86
```

Table 4.3: Details on an object of class `SAEforest`.

Object of class `SAEforest`

Component	Short description
`MERFmodel`	The `MERFmodel` object comprises information on the model fit, details on the algorithm and variance components.
`Indicators`	Element comprising area-level identifiers and estimates.
`MSE_estimates`	Includes area-level identifiers and uncertainty estimates if requested and `NULL` otherwise.
`AdjustedSD`	If MSE results are requested residual variance proposed by Mendez and Lohr (2011) is reported and `NULL` otherwise.
`NrCovar`	Exists only if `meanOnly = TRUE`. Set to `NULL` except `aggData = TRUE` for which it includes a list of variable names of covariates used for the calculation of calibration weights. See Krennmair et al. (2022b) for details.

Details on `MERFmodel`

Component	Short description
`Forest`	Random forest of class `ranger` modelling fixed effects of the model.
`EffectModel`	Model of random effects of class `merMod` capturing structural components of MERFs.
`RandomEffects`	List element containing the values of random intercepts from `EffectModel`.
`RanEffSD`	Standard deviation of random intercepts.
`ErrorSD`	Standard deviation of unit-level errors.
`VarianceCovariance`	VarCorr matrix from `EffectModel`.
`LogLik`	Vector of log-likelihood of the MERF algorithm.
`IterationsUsed`	Iterations used until convergence of the MERF algorithm is reached.
`OOBresiduals`	Vector of OOB-residuals.
`Random`	Character specifying the random intercept in the random effects model.
`ErrorTolerance`	Value monitoring the MERF algorithm's convergence.
`initialRandomEffects`	Vector of initial specification of random effects.
`MaxIterations`	Value specifying the maximal amount of iterations for the MERF algorithm.
`call`	The summarized function call for the object.
`data_specs`	Data characteristics such as domain-specific sample sizes or number of out-of-sample areas.
`data`	The survey sample data.

The `summary` output provides preliminary insights into SAE characteristics such as domain-specific sample sizes, information on sampled and unsampled domains and the total amount of observations. In this example, we face domain-specific sample sizes with a median of 22.5 households, motivating the use of model-based SAE. Moreover, for 24 out of 94 domains, no direct estimates are obtainable. The second essential insight from the output reports model-specific metrics. Starting with the random forest part, we find values such as tuning parameters and R^2 on fixed effects. The R^2 of around 0.62 substantiates the model's predictive capability. The information on the fit of the structural component of our MERF model describes the variance for the area-level random intercept and the individual residuals as well as the intra-class-correlation coefficient (ICC). The ICC of about 0.29 justifies the need for an area-level random effect. The last block of our summary-output highlights convergence properties of the MERF algorithm, such as the amount of needed iterations and the monitored level of likelihood.

As discussed in Section 4.2, the MERF model is a composite model of a random forest and a structural model. This structure is not only mirrored in the output of `summary`, but also within each fitted model object. Thus, users can address elements directly from the fitted model object and use the generic functions from **ranger** (Wright and Ziegler, 2017) and **lme4**

(Bates et al., 2015), respectively. Corresponding objects are stored in `ForestModel` and `Effectmodel`. Especially for objects of class `merMod` (Bates et al., 2015), there exist advantageous generics to extract model components. The following functions are directly applicable: `getData, VarCorr, sigma, residuals, ranef, fixef`. For instance, `ranef` obtains random effects and `VarCorr` directly accesses the variance-covariance matrix:

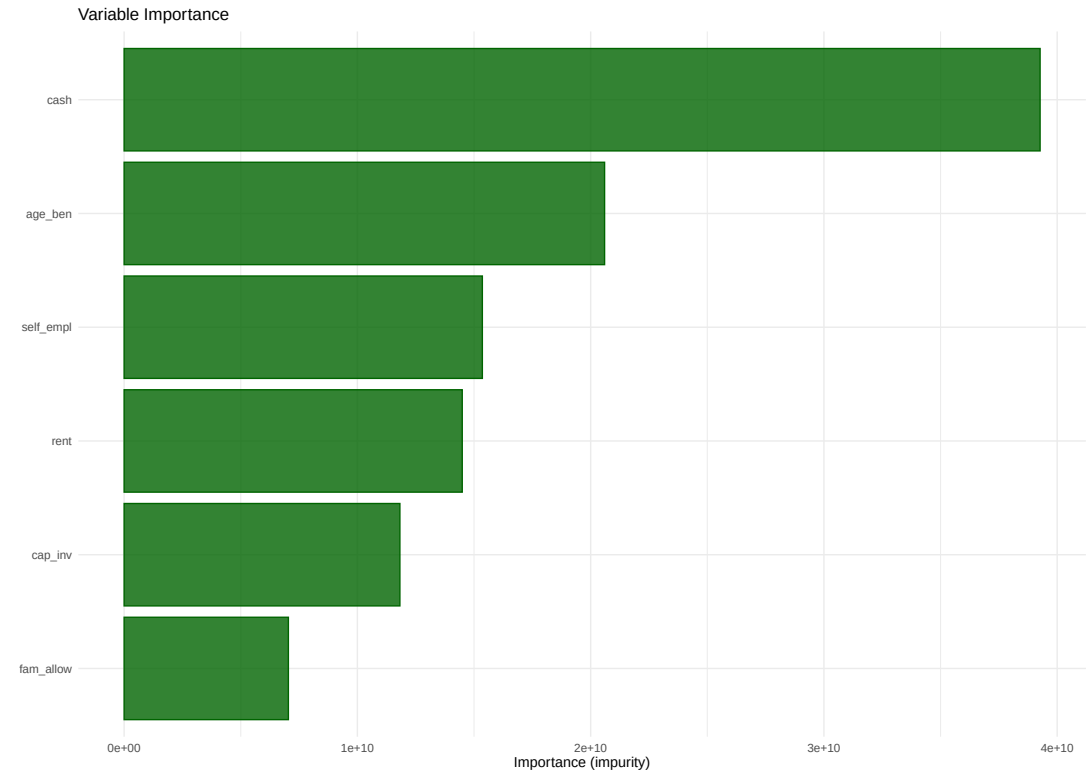
```
R> ranef(MERFmodel1)
R> VarCorr(MERFmodel1)
```

An major complement of summaries and descriptive statistics are diagnostic plots. The generic `plot` function in the package **SAEforest**, produces random forest specific diagnostic tools, like variable importance plots (vip) and partial dependence plots (dpd). A variable importance plot ranks the importance of predictive covariates in the estimation process of the model. Figure 4.1 reports the mean decrease in impurity (variance) calculated for each predictor as the sum over the number of splits across all trees that include the predictor. For the variable importance plot, arguments are passed internally to the function `vip` (Greenwell et al., 2020). The additional partial dependence plot (pdp) depicts the estimated marginal effect for a given number of influential covariates on the target variable. The pdp plot is produced using the package **pdp** (Greenwell, 2017).

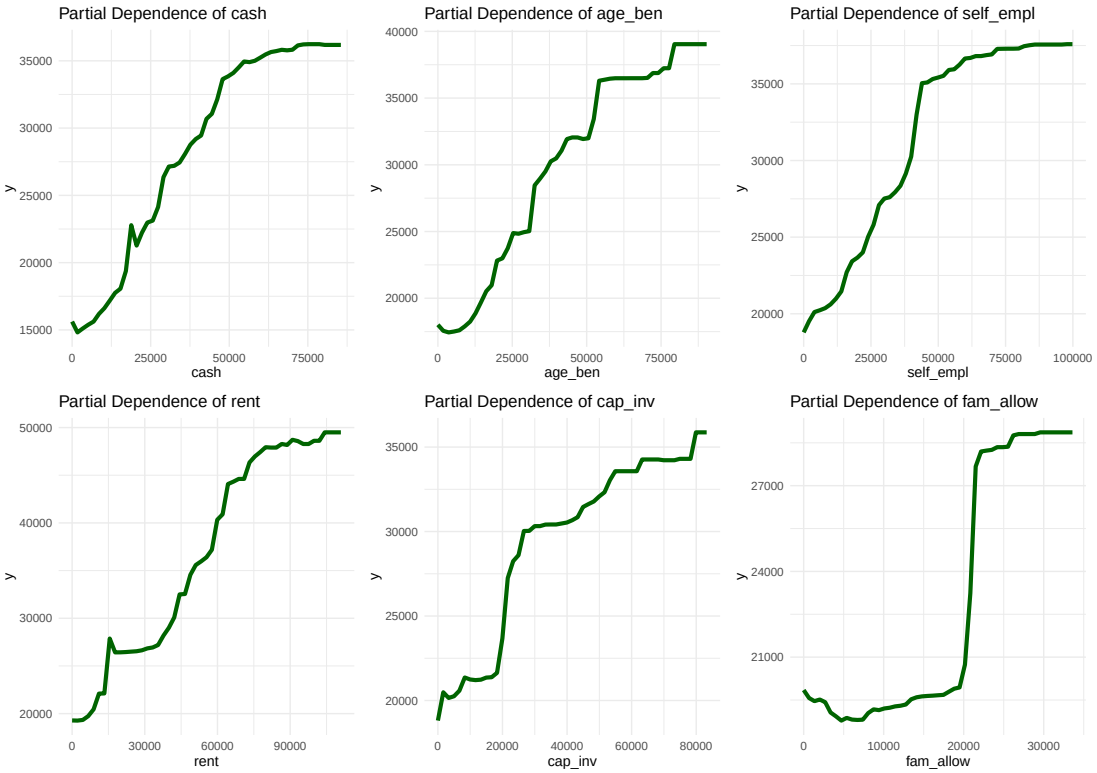
The function `plot` offers several options of customization: most importantly, users can decide whether they want both plots or just the vip plot by specifying `pdp_plot = FALSE`. The plotting engine is **ggplot2** (Wickham, 2016) and several graphical arguments, such as colours or themes can be directly specified. Additionally, the method function `plot` provides the possibility to export a list including requested plots, which allows for modifications based on the additivity of layers for `ggplot`-objects.

```
R> plot(MERFmodel1, num_features = 6, col = "darkgreen",
+   fill = "darkgreen", alpha = 0.8,horizontal = TRUE,
+   gg_theme = theme_minimal(), lsize = 1.5, lty = "solid",
+   grid_row = 2, out_list = FALSE,
+   pdp_plot = TRUE)
```

Figure 4.1 shows the first plot on the fitted object `MERFmodel1`. Most influential variables in the estimation process of fixed effects are net cash income (cash), age-related benefits (age_ben), whether a person is self-employed (self_empl), obtains income from rent (rent), profits from capital investment (cap_inv) or receives family related allowances (fam_allow). Importance plots do not allow for inferences on predictive relations between our target variable of equivalized household income and the covariates. A scrutiny of the pdp plot in Figure 4.1 highlights potential non-linear relations for instance for cash, where the average marginal effect flattens with cash values over 50000. A similar pattern is observable for self-employed income. Another non-linear peculiarity is the discontinuity for fam_allow around 20000.



(a) Variable importance plot (vip).



(b) Partial dependence plots (pdp) for 6 most influential variables.

Figure 4.1: Output from function `plot`.

4.4.3 Model-tuning and important parameters

Random forests are nonparametric procedures, which performance depends on tuning parameters. Function `tune_parameters` assists in fine-tuning of parameters for the implemented MERF method. Essentially, this function is a modified wrapper for `train` from the package **caret** (Kuhn, 2022), treating MERFs as a custom method. Tuning can be performed on the following four parameters: `num.trees` (the number of trees for a forest), `mtry` (number of variables as split candidates at in each node), `min.node.size` (minimal individual node size) and `splitrule` (general splitting rule of individual trees).

Necessary inputs for `tune_parameters` are control parameters for function `train` from package **caret** (Kuhn, 2022), such as the type of cross validation (`method = "repeatedcv"`), the number of folds (`number = 5`), and corresponding repetitions (`repeats = 1`). Moreover, the input of potential tuning parameters must be defined by a grid of parametrization candidates. Data-specific inputs, such as the defined target variable, covariates and the survey dataset resemble the input for the wrapper function `SAEforest_model` discussed in Section 4.4.1.

```
R> fitControl <- caret::trainControl(method = "repeatedcv",
+         number = 5, repeats = 3)

# Define a tuning-grid
R> merfGrid <- expand.grid(num.trees = 500, mtry = c(3,7,9),
+         min.node.size = c(10), splitrule = "variance")

R> tune_parameters(Y = income, X = X_covar, data = eusilcA_smp,
+         dName = "district", trControl = fitControl, tuneGrid =
+         merfGrid, plot_res = FALSE)

1945 samples
15 predictor
No pre-processing
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 1557, 1557, 1556, 1556, 1554, ...
Resampling results across tuning parameters:
mtry  RMSE      Rsquared   MAE
3     5769.200  0.7126250  3832.716
7     5496.742  0.7333739  3565.051
9     5514.225  0.7306313  3556.285
Tuning parameter 'num.trees' was held constant at value of 500
Tuning parameter 'min.node.size' held constant at value of 10
Tuning parameter 'splitrule' held constant at value of variance
RMSE used to select the optimal model using the smallest value.
Final values used for the model were num.trees = 500, mtry = 7,
min.node.size = 10 and splitrule = variance.
```

The output of `tune_parameters` coincides with output from `train` in the package **caret** (Kuhn, 2022). Users can specify whether the summarized information should be accompanied by visualized diagnostics based on **ggplot2** (Wickham, 2016). Most important metrics for fine-tuning decisions are cross-validated results of the RMSE, MAE or the conditional R^2. Following the default specifcation using RMSE as most important criterion for regression, the optimal tuning parameter on the number of randomized split candidates at each node (`mtry`) is 7.

4.4.4 Mapping of results and presentation of indicators

The previous functions focussed on the estimation of indicators and the diagnosis of model quality as well as improvements using optimized tuning parameters. Equally important to the package **SAEforest**, however, is the clear and intuitive presentation of results. Function `summarize_indicators` reports point and MSE estimates as well as calculated coefficients of variation (CV) from a fitted SAEforest object. The CV is an established indicator for national statistical offices to assess associated uncertainty and quality of estimates and is defined as:

$$\text{CV}(\hat{\delta}_i) = \frac{\sqrt{\widehat{\text{MSE}}(\hat{\delta}_i)}}{\hat{\delta}_i}.$$

Users can optionally include a character vector specifying indicators to be reported, referring to all calculated indicators (`all`); each default indicator's name (`Mean, Quant10, Quant25, Median, Quant75, Quant90, Gini, Hcr, Pgap, Qsr` or the function name/s of `custom_indicator`/s) or a vector of multiple indicator names. If the object is estimated by `SAEforest_model` under option `meanOnly = TRUE`, all indicator arguments are ignored and only the `Mean` is returned.

The output object of class `summarize_indicators.SAEforest` allows for generic functions for `data.frames` such as `head, tail, as.matrix, as.data.frame` and `subset`. In the following example, we provide a summary on the Mean, Gini and our customized indicator, identifying the area-level maximum income and respective CVs.

```
R> head(summarize_indicators(MERFmodel2, MSE = FALSE, CV =TRUE,
+        indicator = c("Mean", "Gini", "my_max")))
```

```
district         Mean  Mean_CV Gini Gini_CV my_max my_max_CV
Amstetten      14249.76 0.055 0.248 0.070 56579.45 0.334
Baden          22648.20 0.030 0.177 0.066 69621.40 0.296
Bludenz        12411.98 0.096 0.277 0.091 45723.53 0.456
Braunau am Inn 12046.12 0.069 0.277 0.070 53530.96 0.386
Bregenz        32554.19 0.031 0.156 0.115 77513.46 0.236
```

Revealing spatial patterns of inequality and poverty necessitates the presentation of results with maps. Function `map_indicators` visualizes estimates from a fitted model object of class `SAEforest` on a specified map. Essential inputs for `map_indicators` are the fitted model object, the `map_object` of class `SpatialPolygonsDataFrame` (Bivand et al., 2013) and the domain-level identifier from the `map_object`. For differing

area-level identifiers between the model object of class `SAEforest` and the `map_object`, `map_tab` provides a possibility to enter a `data.frame` linking areas effectively. Comparably to `summarize_indicators`, users can choose specific indicators and whether MSE or CV results should be mapped. For further details, we refer to the help page of function `map_indicators` or Bivand et al. (2013) for a concise overview on the handling of spatial data in R.

Emphasis lies on the flexibility to customize and adapt produced maps. Users can choose colours and themes of the plot based on the plot engine **ggplot2** (Wickham, 2016) and export a list of `ggplot`-elements for further customization if `return_plot = TRUE`. Additionally, users can export a fortified data frame comprising map data and the chosen indicators to produce customized maps using preferred alternative mapping and plotting procedures.

Continuing on our example, we load the shape file on 94 Austrian districts and map results from the fitted object `MERFmodel2` for the Mean and the Gini.

```
R> data("shape_Aut")
R> map_indicators(object = MERFmodel2, MSE = FALSE, CV = TRUE,
+         map_obj = shape_Aut, indicator = c("Mean", "Gini"),
+         map_dom_id = "PB")
```

The map of mean equivalized household income shown by Figure 4.2 indicates differences across Austrian districts, where "Mödling" reports the highest value, which is in accordance to official statistics of income in Austria (Statistik Austria, 2021). Also inequality measured by the Gini is not equally distributed ranging from 0.141 ("Urfahr-Umgebung") to a maximum of 0.301 ("Zell am See"). The majority of CVs for domain-specific values of mean and Gini estimates lies below the 20% threshold, which meets the reliability criterion of Eurostat (2019).

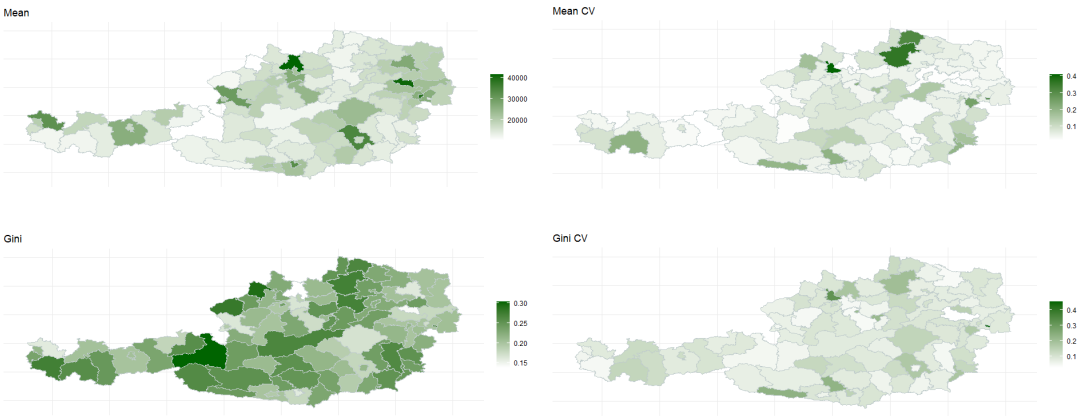


Figure 4.2: District-level estimates for Mean and Gini-coefficient including CVs mapped on Austrian territory. Resulting plots are produced from function `map_indicators`.

4.5 Discussion and outlook

This package aims to bridge concepts of machine learning methods and 'traditional' perspectives of SAE. From a methodological perspective, the estimation of point and uncertainty estimates for domain-level indicators is performed under unit-level and aggregated covariates and dependency structures of observations are modelled using a semi-parametric framework of MERFs. Benefits of random forests align with the proclaimed focus on robustification of SAE-models against model-failure (e.g. providing insurances against model-misspecification, valid variable selection including complex and potentially non-linear interactions between covariates and the effective handling of outliers) (Jiang and Rao, 2020). Moreover, random forests handle high-dimensional ($p > n$) datasets enabling additional perspectives on research concerning Big Data sources (Marchetti et al., 2015; Schmid et al., 2017).

The package **SAEforest** adds valuable insights and advantages to the existing repertoire of SAE methods and yet remains within the methodological tradition of SAE. This includes efforts to provide solutions within the context of domain-level indicators, dependent data structures and in the broader context of survey methodology. We acknowledge that compared to LMMs, benefits of flexibility serve at cost of explainability and attribution, however, this is mitigated by the package's emphasis on informative summary diagnostics and plots (e.g. vip and pdp plots). In addition, the package functionality is characterized by an intuitive workflow and functions to facilitate the visualization of geospatial data. Future versions of the package will ideally include a generalization of our framework to binary and count data. Additionally, the extension towards other machine learning approaches, such as Support Vector Machines, Gradient Boosting and Bayesian Additive Regression Trees is a thought-provoking goal for further research.

Appendix D

D.1 Explanation of variables

Table D.1: Details on the predictive covariates in the survey and population-level datasets.

Variable	Explanation
`eqIncome`	numeric; a simplified version of the equivalized household income. Only available in the survey sample.
`eqsize`	numeric; the equivalized household size according to the modified OECD scale.
`gender`	factor; the person's gender (levels: male and female).
`cash`	numeric; employee cash or near cash income (net).
`self_empl`	numeric; cash benefits or losses from self-employment (net).
`unempl_ben`	numeric; unemployment benefits (net).
`age_ben`	numeric; old-age benefits (net).
`surv_ben`	numeric; survivor's benefits (net).
`sick_ben`	numeric; sickness benefits (net).
`dis_ben`	numeric; disability benefits (net).
`rent`	numeric; income from rental of a property or land (net).
`fam_allow`	numeric; family/children related allowances (net).
`house_allow`	numeric; housing allowances (net).
`cap_inv`	numeric; interest, dividends, profit from capital investments in unincorporated business (net).
`tax_adj`	numeric; repayments/receipts for tax adjustment (net).
`state`	factor; state (nine levels).
`district`	factor; districts (94 levels).
`weight`	numeric; constant weight.

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.

Anderson, W., S. Guikema, B. Zaitchik, and W. Pan (2014). Methods for estimating population density in data-limited areas: Evaluating regression and tree-based models in Peru. PLoS One 9(7).

Asian Development Bank (2021). Practical guidebook on data disaggregation for the sustainable development goals. Policy research report.

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

Atkinson, A. B. (1987). On the measurement of poverty. Econometrica 55(4), 749–764.

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.

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

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

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

Bilton, P., G. Jones, S. Ganesh, and S. Haslett (2020). Regression trees for poverty mapping. Australian & New Zealand Journal of Statistics 62(4), 426–443.

Bivand, R. S., E. Pebesma, and V. Gomez-Rubio (2013). Applied Spatial Data Analysis with R (2nd ed.). New York: Springer.

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

Breckling, J. and R. Chambers (1988). M-quantiles. Biometrika 75(4), 761–771.

Breidenbach, J. (2018). **JoSAE**: Unit-Level and Area-Level Small Area Estimation. R package version 0.3.0.

Breiman, L. (1984). Classification and Regression Trees (1st ed.). New York: Routledge.

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

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

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science 16(3), 199–231.

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.

Capitaine, L. (2020). **LongituRF**: Random Forests for Longitudinal Data. R package version 0.9.

Capitaine, L., R. Genuer, and R. Thiébaut (2021). Random forests for high-dimensional longitudinal data. Statistical Methods in Medical Research 30(1), 166–184.

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.

Chambers, R. and N. Tzavidis (2006, 06). M-quantile models for small area estimation. Biometrika 93(2), 255–268.

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.

Cowell, F. A. (2011). Measuring Inequality (3rd ed.). New York: Oxford University Press.

Dagdoug, M., C. Goga, and D. Haziza (2021). Model-assisted estimation through random forests in finite population sampling. Journal of the American Statistical Association, 1–18.

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

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.

Dietz, G. (2012). Diversity regimes beyond multiculturalism? A reflexive ethnography of intercultural higher education in Veracruz, Mexico. Latin American and Caribbean Ethnic Studies 7(2), 173–200.

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

Efron, B. (2020). Prediction, estimation, and attribution. Journal of the American Statistical Association 115(530), 636–655.

Efron, B. and T. Hastie (2016). Computer Age Statistical Inference. Cambridge University Press.

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

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.

Eurostat (2004). Common Cross-Sectional EU Indicators Based on EU-SILC; the Gender Pay Gap, Chapter EU-SILC 131-rev/04, Unit D-2: Living conditions and social protection, Directorate D: Single Market, Employment and Social statistics. Luxembourg: Eurostat.

Eurostat (2019). DataCollection: precision level DCF. Eurostat, Luxembourg. (Available from `https://datacollection.jrc.ec.europa.eu/wordef/precision-level-dcf`). [accessed: 06.2019].

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.

Feng, X., X. He, and J. Hu (2011). Wild bootstrap for quantile regression. Biometrika 98(4), 995–999.

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

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.

Ghosh, M., J. Myung, and F. Moura (2016). **robustsae**: Robust Bayesian Small Area Estimation. R package version 0.1.0.

Gini, C. (1912). Variabilita e Mutabilita. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche. Bologna: P. Cuppini.

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.

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.

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

Greenwell, B. M., B. Boehmke, and B. Gray (2020). Variable importance plots: An introduction to the **vip** package. The R Journal 12(1), 343–366.

Gutierrez, J. P., M. Agudelo-Botero, S. Garcia-Saiso, C. Zepeda-Tena, C. A. Davila-Cervantes, M. C. Gonzalez-Robledo, N. Fullman, C. Razo, B. Hernández-Prado, G. Martínez, S. Barquera, and R. Lozano (2020). Advances and challenges on the path toward the sdgs: subnational inequalities in Mexico, 1990–2017. BMJ Global Health 5(10).

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 (2011). Mixed effects regression trees for clustered data. Statistics & Probability Letters 81(4), 451–459.

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.

Jiang, J. and J. S. Rao (2020). Robust small area estimation: An overview. Annual Review of Statistics and its Application 7(1), 337–360.

Kilic, T., U. Serajuddin, H. Uematsu, and N. Yoshida (2017). Costing household surveys for monitoring progress toward ending extreme poverty and boosting shared prosperity. World Bank Policy Research Working Paper (7951).

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.

Krennmair, P. (2022). **SAEforest**: Mixed Effect Random Forests for Small Area Estimation. R package version 1.0.0.

Krennmair, P. and T. Schmid (2022). Flexible domain prediction using mixed effects random forests. Journal of Royal Statistical Society: Series C (Applied Statistics) 71(5), 1865–1894.

Krennmair, P., T. Schmid, and N. Tzavidis (2022). The estimation of poverty indicators using mixed effects random forests: Case study for the Mexican state of Veracruz. Workig Paper.

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.

Kuhn, M. (2022). **caret**: Classification and Regression Training. R package version 6.0-92.

Lambert, F. and H. Park (2019). Income inequality and government transfers in Mexico. IMF Working Papers (148).

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

Li, Z. R., B. D. Martin, Y. Hsiao, J. Godwin, J. Paige, J. Wakefield, S. J. Clark, G.-A. Fuglstad, and A. Riebler (2021). **SUMMER**: Small-Area-Estimation Unit/Area Models and Methods for Estimation in R. R package version 1.2.0.

Liaw, A. and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

Lu, Y., N. Nakicenovic, M. Visbeck, and A.-S. Stevance (2015). Policy: Five priorities for the un sustainable development goals. Nature 520(7548), 432–433.

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 N. Tzavidis (2021). Robust estimation of the theil index and the gini coefficient for small areas. Journal of Official Statistics 37(4), 955–979.

Marchetti, S., N. Tzavidis, and M. Pratesi (2012). Non-parametric bootstrap mean squared error estimation for m-quantile estimators of small area averages, quantiles and poverty indicators. Computational Statistics & Data Analysis 56(10), 2889–2902.

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.

Marino, M. F., N. Tzavidis, and M. Alfo (2018). Mixed hidden markov quantile regression models for longitudinal data with possibly incomplete sequences. Statistical Methods in Medical Research 27(7), 2231–2246.

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

Mendez, G. (2008). Tree-Based Mehtods to Model Dependent Data. Ph. D. thesis, Arizona State University.

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

Molina, I. and Y. Marhuenda (2015, jun). **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.

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

Neufeld, A. and B. Heggeseth (2019). **splinetree**: Longitudinal Regression Trees and Forests. R package version 0.2.0.

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.

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].

OECD (2021). OECD social and welfare statistics (database). OECD, Paris. doi: https://doi.org/10.1787/data-00654-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.

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

Owen, A. (2001). Empirical likelihood. New York: Chapman and Hall.

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

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.

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

Rao, J. N. K. and I. Molina (2015). Small Area Estimation (2nd ed.). New Jersey: Wiley: Wiley series in survey methodology.

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

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.

Sachs, J. D., G. Schmidt-Traub, M. Mazzucato, D. Messner, N. Nakicenovic, and J. Rockström (2019). Six transformations to achieve the sustainable development goals. Nature Sustainability 2(9), 805–814.

Saha, A., S. Basu, and A. Datta (2021). **RandomForestsGLS**: Random Forests for Dependent Data. R package version 0.1.3.

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

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

Scornet, E., G. Biau, and J.-P. Vert (2015). Consistency of random forests. The Annals of Statistics 43(4), 1716–1741.

Sela, R. J. and J. S. Simonoff (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. Machine Learning 86(2), 169–207.

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

Singleton, A., A. Alexiou, and R. Savani (2020). Mapping the geodemographics of digital inequality in great britain: An integration of machine learning into small area estimation. Computers, Environment and Urban Systems 82, 101486.

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

Smits, J. and I. Permanyer (2019). The subnational human development database. Scientific Database 6(190038).

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.

Statistik Austria (2021). Statistik der Lohnsteuer 2020. technical report. Statistik Austria, Vienna. https://www.statistik.at/fileadmin/publications/Statistik_der_Lohnsteuer_2020. [accessed: 05.2022].

Statistisches Bundesamt (2015). Zensus 2011 Methoden und Verfahren. Statistisches Bundesamt, Wiesbaden. https://www.zensus2011.de/SharedDocs/Downloads/DE/Publikationen/Aufsaetze_Archiv/2015_06_MethodenUndVerfahren.pdf?__blob=publicationFile&v=6. [accessed: 12.2020].

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.

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.

United Nations (2014). A world that counts: Mobilising the data revolution for sustainable development. Research report, United Nations, New York.

United Nations (2015). Transforming our world: The 2030 agenda for sustainable development. Resolution A/RES/70/1, United Nations, New York.

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

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.

Wang, J. and L. S. Chen (2016). **MixRF**: A Random-Forest-Based Approach for Imputing Clustered Incomplete Data. R package version 1.0.

Wardrop, N., W. Jochem, T. Bird, H. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman, and A. Tatem (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. Proceedings of the National Academy of Sciences 115(14), 3529–3537.

Warnholz, S. (2018). **saeRobust**: Robust Small Area Estimation. R package version 0.2.0.

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

Winham, S. J., R. R. Freimuth, and J. M. Biernacka (2013). A weighted random forests approach to improve predictive performance. Statistical Analysis and Data Mining 6(6), 496–505.

Wood, S. (2017). Generalized Additive Models: An Introduction with R (2nd ed.). New York: Chapman and Hall/CRC.

World Bank Group (2015). A measured approach to ending poverty and boosting shared prosperity: Concepts, data, and the twin goals. Policy research report, The World Bank Group, Washington DC.

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.

Wu, H. and J.-T. Zhang (2006). Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches. New Jersey: John Wiley & Sons.

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

Summaries

Abstracts in English

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

This paper promotes the use of random forests as versatile tools for estimating spatially disaggregated indicators in the presence of small area-specific sample sizes. Small area estimators are predominantly conceptualized within the regression-setting and rely on linear mixed models to account for the hierarchical structure of the survey data. In contrast, machine learning methods offer non-linear and non-parametric alternatives, combining excellent predictive performance and a reduced risk of model-misspecification. Mixed effects random forests combine advantages of regression forests with the ability to model hierarchical dependencies. This paper provides a coherent framework based on mixed effects random forests for estimating small area averages and proposes a non-parametric bootstrap estimator for assessing the uncertainty of the estimates. We illustrate advantages of our proposed methodology using Mexican income-data from the state Nuevo León. Finally, the methodology is evaluated in model-based and design-based simulations comparing the proposed methodology to traditional regression-based approaches for estimating small area averages.
Keywords: official statistics, small area estimation, mean squared error, tree-based methods

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 estima-

tor 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: The estimation of poverty indicators using mixed effects random forests: case study for the Mexican state of Veracruz

Mapping and analysing the spatial concentration of poverty is imperative for evidence-based policies to translate into inclusive and sustainable actions. The use of national sample surveys to obtain detailed and reliable estimates for poverty indicators on disaggregated geographical and other domains (e.g. demographic groups) imposes a methodological challenge. Small Area Estimation is a collective term for (model-based) procedures, which combine survey data with existing auxiliary information (e.g. census or administrative data) using predictive models to estimate domain-specific statistical indicators. We propose the use of mixed effects random forests as flexible, robust, and reliable method to produce domain-specific cumulative distribution functions from which (non-linear) poverty estimators can be obtained. This paper is driven by our aim to inform a transparent and steady discussion on current methodological improvements for Small Area Estimation, such as the use of (tree-based) machine learning methods and their contribution to recent requirements for poverty assessment. We evaluate proposed point and uncertainty estimators in a design-based simulation and focus on a case study uncovering spatial patterns of poverty for the Mexican state of Veracruz.
Keywords: official statistics; small area estimation; mean squared error; mixed models, random forest

Abstract: The R package SAEforest

The R package **SAEforest** promotes the use of Mixed Effects Random Forests (MERFs) for applications of Small Area Estimation. The package effectively combines functions for the estimation of spatially disaggregated linear and non-linear indicators using survey sample data. Models increase the precision of direct estimates from survey data, combining unit-level or aggregated covariate information from auxiliary data. Included procedures facilitate the estimation of domain-level economic and inequality metrics and assess associated uncertainty. The package provides procedures to simplify the analysis of model performance of MERFs and enables the visualization of predictive relations from covariates. Additionally, the package includes a function for fine-tuning of required hyper-parameters. General emphasis lies on straightforward interpretation and mapping of results.
Keywords:official statistics, mixed effects random forests, small area estimation, poverty mapping

Kurzzusammenfassungen auf Deutsch

Zusammenfassung: Flexible Schätzung von Mittelwerten für kleine Stichprobenumfänge mit Mixed-Effects-Random-Forests

In diesem Beitrag wird die Verwendung von Random-Forests als vielseitiges Instrument zur Schätzung räumlich disaggregierter Indikatoren bei kleinen Stichprobengrößen vorgestellt. Schätzer für diese sogenannten Small-Areas werden vorwiegend durch linear gemischte Regressionsmodelle erzeugt, um die hierarchische Struktur der Erhebungsdaten zu berücksichtigen. Im Gegensatz dazu bieten Methoden des maschinellen Lernens nicht-lineare und nicht-parametrische Alternativen, die eine hervorragende Prädiktion mit geringerem Risiko von Modellfehlspezifizierungen kombinieren. Mixed-Effects-Random-Forests kombinieren die Vorteile von baumbasierten prädiktiven Algorithmen mit der Fähigkeit hierarchische Abhängigkeiten zu modellieren. In diesem Beitrag wird ein methodologischer Rahmen für die Grundlage von Random Forests für die Schätzung von Durchschnittswerten für kleine Stichprobenbereiche geschaffen. Darüber hinaus wird ein nicht-parametrischer Bootstrap-Schätzer für die Bewertung der Unsicherheit der Schätzungen vorgeschlagen. Wir veranschaulichen die Vorteile unserer vorgeschlagenen Methodik anhand mexikanischer Einkommensdaten aus dem Bundesstaat Nuevo León. Die Methodik wird in modellbasierten und designbasierten Simulationen evaluiert und traditionellen, regressionsbasierten Verfahren gegenübergestellt.
Schlüsselwörter: Amtliche Statistik, Small-Area-Schätzung, mittlerer quadratischer Fehler, baumbasierte Methoden

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: Die Schätzung von Armutsindikatoren unter Verwendung von Mixed-Effects-Random-Forests: Fallstudie für den mexikanischen Bundesstaat Veracruz

Die Visualisierung und Analyse der räumlichen Konzentration von Armut ist für die Realisierung von langfristigen und nachhaltigen Politikmaßnahmen von Bedeutung. Die Verwendung nationaler Stichprobenerhebungen für detaillierte und zuverlässige Schätzungen von (räumlich) disaggregierten Armutsindikatoren stellt eine methodische Herausforderung dar. Small-Area-Estimation ist ein Sammelbegriff für (modellgestützte) Verfahren, bei denen Erhebungsdaten mit vorhandenen Zusatzinformationen (z. B. Zensus- oder Verwaltungsdaten) unter Verwendung von Vorhersagemodellen kombiniert werden, um statistische Indikatoren für kleine Stichprobenbereiche zu schätzen. Dieser Beitrag empfiehlt die Verwendung von sogenannten Mixed-Effects-Random-Forests als flexible, robuste und zuverlässige Methoden zur Erstellung kumulativer Verteilungsfunktionen für kleine Stichprobenumfänge. Auf Basis dieser geschätzten kumulativen Verteilungsfunktionen können (nicht-lineare) Armutsschätzer gewonnen werden. Ein weiteres Ziel dieser Arbeit ist es, eine transparente und kontinuierliche Diskussion über aktuelle methodische Entwicklungen im Bereich von Small-Area-Estimation zu eröffnen, wie z.B. die Verwendung von Methoden des maschinellen Lernens und ihre Beiträge zur Armutsforschung. Wir evaluieren vorgeschlagene Punkt- und Unsicherheitsschätzer in einer designbasierten Simulation und konzentrieren uns auf eine Fallstudie zur Ermittlung räumlicher Armut für den mexikanischen Bundesstaat Veracruz.
Schlüsselwörter: Amtliche Statistik; Small-Area-Schätzung; mittlerer quadratischer Fehler; gemischte Modelle, Random Forest

Zusammenfassung: Das R-Paket SAEforest

Das R-Paket **SAEforest** erleichtert die Verwendung von Mixed-Effects-Random-Forests für Anwendungen (räumlich) disaggregierter Schätzungen (i.e. Small-Area-Estimation). Das Paket kombiniert Funktionen für die Schätzung (nicht-)linearer Indikatoren unter Verwendung von Erhebungsstichprobendaten. Die Modelle erhöhen die Präzision direkter Schätzungen aus Umfragedaten, indem sie Informationen über Kovariaten auf individualer oder aggregierte Ebene aus verfügbaren alternativen Hilfsdaten kombinieren. Enthaltene Verfahren erleichtern die Schätzung von Wirtschafts- und Ungleichheitsmaßzahlen für räumliche oder demografischen Gruppen und bewerten die verbundene Unsicherheit der Schätzungen. Das Paket bietet Methoden zur einfachen Darstellung der Modeleigenschaften von Mixed-Effects-Random-Forests und ermöglicht die Visualisierung der Vorhersagebeziehungen der Kovariaten. Zusätzlich enthält das Paket eine Funktion zur Feinabstimmung der erforderlichen Hyperparameter. Der Schwerpunkt des Pakets liegt auf einer einfachen Anwendung von Mixed-Effects-Random-Forests sowie der Interpretation und Abbildung der Ergebnisse.
Schlüsselwörter: Amtliche Statistik, Mixed Effects Random Forests, Small-Area-Schätzung, Armutskartierung

Declaration of Authorship

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

Berlin, October 14, 2022

_____

Patrick Krennmair
October 14, 2022