Thesis submitted in fulfilment of the requirements for the degree

**Dr. rer. pol.**

on the topic

# Generalized Tree-Based Machine Learning Methods with Applications to Small Area Estimation

to the

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

submitted by

Nicolas Frink
born in Berlin, Germany


Bamberg 2025

Cumulative dissertation
Nicolas Frink, *Generalized Tree-Based Machine Learning Methods with Applications to Small Area Estimation*

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

First reviewer: Prof. Dr. Timo Schmid
Second reviewer: Prof. Dr. Ulrich Rendtel
Third reviewer: Prof. Dr. Nikos Tzavidis
Date of defense: 25.02.2025

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

Erstgutachter: Prof. Dr. Timo Schmid
Zweitgutachter: Prof. Dr. Ulrich Rendtel
Drittgutachter: Prof. Dr. Nikos Tzavidis
Tag der mündlichen Prüfung: 25.02.2025




URN: urn:nbn:de:bvb:473-irb-1076605
DOI: https://doi.org/10.20378/irb-107660

## Acknowledgements

First of all, I wish to express my profound gratitude to my supervisor, Prof. Dr. Timo Schmid, who has continually encouraged and supported me throughout my studies. His introductory statistics course in the summer term of 2015 not only showed me the various possibilities in the field of statistics as an undergraduate, but also sparked my interest in the topic in numerous other courses. This brings us full circle, as he supervised my first empirical work and now my most important work to date, this dissertation. He assisted me at every stage of my thesis and has been a constant source of professional and personal guidance. Thank you for everything you have done for me.

I am also very grateful to Prof. Dr. Ulrich Rendtel. His insightful discussions have made me a better statistician and significantly improved this thesis. At the same time, he showed that one can be passionate about statistics even after so many years, without losing the joy and fun of it.

Many thanks to Prof. Dr. Nikos Tzavidis for his scientific expertise and intuition. His feedback was always clear and insightful, and his suggestions often led me to see things from a different perspective.

My sincere thanks also go to Prof. Dr. Jan Marcus, who has always supported me and shown me how to best present my research projects. Through our collaboration, I have encountered new and exciting challenges, both professionally and academically, which will undoubtedly benefit me in the future.

I am very thankful to my colleagues at the Chair of Applied Statistics at Freie Universität Berlin, the Statistical consulting unit fu:stat and the Faculty of Social and Economic Sciences at the University of Bamberg for providing me with a pleasant working and learning environment. In particular to Angelika, Felix, Lisa, Lorena, Lukas, Marc B., Marc S., Nora, Patrick, Sylvia, Ulrich and Yeonjoo.

I would like to thank my parents and brother, Jana, Hans-Peter and Maximilian for their permanent support and patience over the years.

Most importantly, I would like to thank my wife, Untera, from the bottom of my heart, who has always accompanied me on this journey and has always supported and encouraged me. I have been able to experience the positive and challenging aspects together with you, and you have always had an open ear for the things that have been on my mind during the creation of this thesis. None of this would have been possible without you.

## Publication List

The publications listed below are the result of the research carried out in this thesis titled, "Generalized Tree-Based Machine Learning Methods with Applications to Small Area Estimation".

1. Frink, N., and Schmid, T. (2024a) **Small area estimation with generalized random forests: Estimating poverty rates in Mexico**, *Working paper*, submitted.

2. Frink, N., and Schmid, T. (2024b) **Small area prediction of counts under machine learning-type mixed models**, *Working paper*, submitted.

3. Frink, N. (2024) **A framework for generalizing mixed effect random forests in R**, *Working paper*, to be submitted.

# Contents

# Introduction

Sample surveys serve as a crucial tool for governments, policymakers, and organizations in making informed decisions. Decision-makers increasingly require information not only for the overall population but also for specific subgroups or areas. These subgroups may include geographic regions like states, municipalities, or school districts, as well as socio-demographic categories such as gender and age groups (Tzavidis et al., 2018). For instance, in the context of development aid, organizations frequently encounter the challenge of acquiring precise data on living conditions in small, remote, or sparsely researched areas (Tarozzi and Deaton, 2009). Similarly, in the field of economics, companies and policymakers often require detailed information on consumer behavior, income, or unemployment in small regions to make informed decisions (Pfeffermann, 2013). In politics, precise data is also essential for planning and implementing social programs and political measures that effectively address the diverse needs of various population groups and regions (Lehtonen and Veijanen, 2009). Nevertheless, conventional surveys are often designed to produce indicators solely at the national or regional level. The costs and time required for comprehensive surveys often render them impractical for smaller geographical areas. Since traditional surveys cannot cover every minor geographical area and small sample sizes are often insufficient, deriving meaningful estimates from survey data alone is usually not feasible. Consequently, the plausibility of these estimates is inherently questionable and, in many instances, they cannot be quantified (Pfeffermann, 2011; Rao and Molina, 2015; Morales et al., 2021).

To counteract this problem, small area estimation (SAE) has become an important tool in the field of social statistics (Elbers et al., 2003). An area is considered to be small if the sample size specific to the area is insufficient to produce estimates with the required precision (Ghosh, 2020). The core idea of SAE methods is to leverage the relationships among related areas and incorporate various sources of information to improve the reliability of domain estimates (Ghosh and Rao, 1994; Jiang and Lahiri, 2006). This is typically achieved through the use of mixed models. These models link the relevant areas either explicitly or at least implicitly (Ghosh, 2020). To achieve this, model-based SAE methods require two types of datasets: a survey dataset and an auxiliary dataset (such as a census dataset or alternative data sources like mobile phone data, (Schmid et al., 2017)). The survey dataset, typically collected at the individual or household level, must include the dependent variable alongside potential independent variables. In contrast, the auxiliary dataset needs to contain only the independent variables. The combination of these datasets allows for more precise and accurate estimation, addressing the limitations of sparse survey data (Rao and Molina, 2015). Recent advancements in model-based SAE methodologies have considerably improved the reliability of indicators

at the micro-level, providing policymakers with actionable insights to address socio-economic disparities (Datta and Ghosh, 2012).

While much of the initial focus in SAE has been on continuous outcomes, there is a growing need to address discrete responses, such as binary or count data, which are common in many socio-economic applications. Discrete response models present unique challenges due to the non-linear nature of the data and the need to account for the distributional characteristics of the responses. Model-based SAE approaches for discrete responses frequently involve generalized linear mixed models (GLMM), which extend linear mixed models to handle non-normal data distributions (Fahrmeir and Tutz, 2001). These models can incorporate random effects to account for the hierarchical structure of the data, enabling more accurate estimation for small areas (Malec et al., 1997; Ghosh et al., 1998; Jiang and Lahiri, 2001). For example, mixed Logistic models are used for binary outcomes, while mixed Poisson or Negative Binomial models are employed for count data (Jiang and Nguyen, 2021). Despite their advantages, traditional GLMMs for discrete responses are parametric models and rely on model assumptions that may not always align with empirical evidence. Violation of these assumptions can lead to biased parameter estimates and incorrect inferences (Jiang and Rao, 2020). In order to avoid the pitfalls of parametric assumptions associated with GLMMs, machine learning techniques offer a viable alternative methodology.

Machine learning methods have established themselves as central technologies in modern data analysis over recent decades. They encompass a variety of algorithms and techniques that enable computers to learn from data and make predictions or decisions without explicit programming. Machine learning is used in many areas, including image and speech recognition, medical diagnostics, financial market analysis and autonomous driving. A key advantage of these methods is their ability to recognize and model complex patterns and relationships in datasets, without relying on explicit model assumptions, which often overwhelm traditional statistical methods (Sarker, 2021). Machine learning methods can essentially be divided into two main categories: supervised and unsupervised learning. In supervised learning, models are trained using labeled datasets, where each input is linked to the desired output. The most common applications of supervised learning include classification and regression tasks. In contrast, unsupervised learning works with unlabeled data and aims to discover hidden patterns or structures in the data (Hastie et al., 2009). Among the various supervised and unsupervised machine learning methods available, tree-based approaches are particularly beneficial in survey research (Kern et al., 2019). Tree-based methods include decision trees (Breiman et al., 1984), bagging (Breiman, 1996), random forests (Breiman, 2001), and gradient boosting machines (Friedman, 2001). These methods are compelling because they make few assumptions, aside from the independence of observations. They simultaneously perform implicit model selection and allow for flexible specification of the relationship between the outcome and the covariates. However, SAE applications typically involve hierarchical data, where observations are often correlated. Ignoring these correlations usually results in less accurate point predictions and inferences.

Therefore, in this thesis I combine the respective advantages of traditional parametric statistical methods and modern non-parametric predictive machine learning methods by introducing the generalized mixed effect random forest (GMERF) in the context of small area estimation.

GMERFs are able to model discrete outcomes, thereby extending the functionality of the mixed effect random forest (MERF, (Hajjem et al., 2014; Krennmair and Schmid, 2022)), which was originally designed for continuous target variables. GMERFs unite the strengths of random forests with the capability to manage dependency structures in survey data. In particular, the GMERF is presented for two different sub-cases: binary and count target variables.

Chapter 1 introduces the GMERF for estimating poverty indicators based on binary outcome variables and associated uncertainties within the SAE context. Compared to existing SAE methods for binary target variables, the GMERF estimator sidesteps model selection issues, effectively captures non-linear interactions and higher-order effects, and manages high-dimensional covariate data even when the number of covariates exceeds the sample size. We evaluate the proposed point and uncertainty estimators using model- and design-based simulations, focusing on a case study that reveals spatial patterns of poverty in the Mexican state of Tlaxcala. Beyond this methodological contribution, Chapter 1 also examines how converting continuous information into binary information affects the accuracy of estimation methods. Our findings demonstrate that, even with limited information, the GMERF can deliver performance comparable to methods that require more detailed continuous data. The first chapter focuses exclusively on binary indicators and does not cover the GMERF for count data.

To address this research gap, Chapter 2 introduces the GMERF for count indicators, along with a MERF designed to handle challenges specific to count outcomes, such as overdispersion. Chapter 2 presents and evaluates three bootstrap methodologies, two non-parametric and one parametric, developed to assess the reliability of point estimators for area-level means. The efficacy of these methodologies is evaluated through model- and design-based simulations and applied to a real-world dataset from the state of Guerrero in Mexico, demonstrating their resilience and practical uses. The results indicate that the MERF, which does not rely on Poisson distribution assumptions to model the mean behavior of count data, is particularly effective in scenarios of severe overdispersion. In contrast, the GMERF performs best under conditions where Poisson distribution assumptions are moderately fulfilled.

In the first two chapters of this thesis, relevant extensions are provided. To make these methods accessible to users, the final chapter focuses on their implementation in statistical software. Chapter 3 bridges the gap between academic theory and practical application by integrating the discussed methodologies for point and uncertainty estimation into a new version of the R package **SAEforest**. Version 2.0.0 incorporates various distributions and link functions, thereby providing a comprehensive solution for a wider array of modeling requirements in small area estimation. Both MERF and GMERF models can now be conveniently estimated using a single function. Furthermore, we have integrated new user-friendly diagnostic plots and introduced an updated tuning function for the components of the generalized random forest model.

# Chapter 1

# Small area estimation with generalized random forests: Estimating poverty rates in Mexico

**Abstract**

Identifying and addressing poverty is challenging in administrative units with limited information on income distribution and well-being. To overcome this obstacle, small area estimation methods have been developed to provide reliable and efficient estimators at disaggregated levels, enabling informed decision-making by policymakers despite the data scarcity. We propose a robust and flexible approach for estimating poverty indicators based on binary response variables within the small area estimation context: the generalized mixed effects random forest. Our method employs machine learning techniques to identify predictive, non-linear relationships from data, while also modeling hierarchical structures. Mean squared error estimation is explored using a parametric bootstrap. From an applied perspective, we examine the impact of information loss due to converting continuous variables into binary variables on the performance of small area estimation methods. We evaluate the proposed point and uncertainty estimates in both model- and design-based simulations. Finally, we apply our method to a case study revealing spatial patterns of poverty in the Mexican state of Tlaxcala.

**Keywords**: Data integration, Generalized mixed models, MSE estimation, Parametric bootstrap

## 1.1 Introduction

Poverty is a primary feature of development challenges, and reducing poverty is the first sustainable development goal (SDG) of the United Nations and one of the World Bank twin goals (United Nations, 2015; World Bank Group, 2015). However, the COVID-19 pandemic has dealt a severe blow to poverty reduction efforts, with poverty rates increasing for the first time in two decades (World Bank Group, 2020). This effect is particularly pronounced in regions with already high poverty rates, such as Latin America, exacerbating the impoverishment of vulnerable populations. As a result, the World Bank's goal of ending poverty by 2030 seems increasingly unachievable without significant targeted policy actions (World Bank Group, 2015). To effectively combat poverty, governments must be able to identify areas with the greatest need for intervention, channeling resources, skills, and innovation to these regions. However,

many areas lack information on income distribution and well-being, making it difficult to identify and address poverty. This lack of information is often due to limited funding for population surveys, particularly for smaller spatially disaggregated areas such as states, districts, or municipalities (Tzavidis et al., 2018; Asian Development Bank, 2021). Small area estimation (SAE) methods were developed to provide reliable and efficient estimators on disaggregated levels that allow policymakers to make informed decisions despite the lack of available information (Rao and Molina, 2015). The head count ratio (HCR) is an established economic indicator representing the percentage of households living below a certain poverty line within a given region. The HCR is based on a binary variable (Peragine et al., 2021), taking a value of 1 if the household income is below the poverty line $t$ and 0 otherwise (Foster et al., 1984). The HCR for an area $i$ is defined as:

$$\text{HCR}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{1}(z_{ij} \leq t),$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, $z_{ij}$ represents the income of household $j$ and $N_i$ is the population size of area $i$. Previous literature on SAE has estimated this binary-based poverty indicator using models that assume continuous input variables, such as household income. Restricted access to continuous unit-level household income data due to data security concerns, high non-response rates, or data collection limitations presents a challenge for SAE practitioners. For example, the National Health Interview Survey (NHIS) (National Center for Health Statistics, 2023) does not provide continuous household income data for public use.

A variety of methods for SAE with binary outcomes have been explored. For instance, Malec et al. (1997) and Nandram et al. (1999) propose a hierarchical Bayes (HB) approach. Alternatively, empirical Bayes (EB) methods, as described by MacGibbon and Tomberlin (1989) and Farrell et al. (1997), offer another approach. From a frequentist perspective, the empirical best predictor (EBP) suggested by Jiang and Lahiri (2001) and Jiang (2003) represents a viable option. Building on this, González-Manteiga et al. (2007) develop methods for estimating the mean squared error (MSE) of small-area predictors under logistic mixed models, specifically using parametric bootstrapping. Hobza and Morales (2016) derive an approximation of the MSE for the EBP and propose a corresponding bias-correction. Moreover, Chambers et al. (2016) introduce M-quantile modeling as a robust alternative for binary outcomes, offering protection against model misspecification and outliers. However, each of these methods assumes that a specified transformation of the expectation function can be expressed as a linear combination of covariates.

Hence, this paper introduces the generalized mixed effects random forest (GMERF) within the methodological tradition of SAE as a novel, flexible, data-driven, and semi-parametric approach for estimating poverty rates based on binary variables. We incorporate the random forest (Breiman, 2001) within the mixed models framework because it exhibits excellent predictive performance, captures relationships directly from the data, and can incorporate higher-order interactions between covariates without requiring explicit model assumptions (Hastie et al., 2009; Biau and Scornet, 2016).

Machine learning methods, in general, have already been applied in SAE (Bilton et al., 2017, 2020). Nonetheless, they exhibit limitations, such as disregarding the correlation among

subpopulations and the incapacity to handle intricate covariance structures. Recently, there has been a growing interest in expanding the application of one particular machine learning technique, specifically the tree-based method, by incorporating it into mixed effects models. This integration allows for the surpassing of the inherent limitations associated with these models. For instance, Krennmair and Schmid (2022) and Krennmair (2022) introduce the concept of mixed effect random forests (MERF) within the methodological framework of SAE. They assume a continuous response variable and substitute the linear combination of covariates in the fixed effects section of a linear mixed model (LMM) with a regression forest. This approach is appropriate for Gaussian response variables but not suitable for classification problems. For binary response variables and individual predictions, different methods exist. For example, Hajjem et al. (2017) propose a generalized mixed effects regression tree (GMERT), while Fontana et al. (2021) present the generalized mixed effects tree (GMET). GMET utilizes the tree leaves as indicator variables instead of the tree predictions used in GMERT. Although existing methods extend the use of simple trees for modeling nested data with non-Gaussian response variables, they do not incorporate tree ensembles. Therefore, Pellagatti et al. (2021) extend the GMET approach (Fontana et al., 2021) and use random forests instead of standard trees in the fixed effects section of the mixed effects model. Although the aforementioned methods combine machine learning and mixed models for estimating binary variables, they mainly focus on individual predictions.

The major methodological contribution of our paper is the extension of the GMERT approach (Hajjem et al., 2017) to a GMERF and incorporating it into the SAE framework for estimating area-level proportions, for instance poverty rates. Furthermore, we propose a parametric mean squared error-bootstrap scheme to asses the uncertainty associated with area-level estimates. In contrast to the predominantly regression-based 'traditional' SAE methods for binary target variables (Jiang and Lahiri, 2001; Jiang, 2003), the GMERF offers protection against model misspecification and captures non-linear relationships from data. This highlights a major drawback of the classical SAE methods, as they rely on model assumptions that barely match real findings when considering data on social and economic inequality. This discrepancy between assumptions and reality can introduce bias in parameter estimates, making MSE estimates unreliable (Jiang and Rao, 2020).

In addition to our methodological contribution, we also aim to discuss the impact of information loss caused by converting continuous input variables into binary variables on the performance of estimation methods. To achieve this, we evaluate the performance of our GMERF against established SAE methods that use continuous household income as an input variable. These methods include the EBP with data-driven transformation (Rojas-Perilla et al., 2020), the MERF for non-linear indicators (Krennmair, 2022), and the Fay-Herriot (FH) model with a logit transformation that accounts for the survey design (Runge, 2023).

The paper is structured as follows: Section 1.2 introduces GMERFs as a method that integrates random forests and hierarchical modeling to account for dependencies between unit-level observations. In Section 1.2.2, we elucidate the construction of area-level proportion estimates and deliberate upon the specific data scenarios in which our method works particularly well. To address the issue of accurately estimating the MSE of the area-level estimates, Section

1.3 proposes a parametric bootstrap scheme. In Section 1.4, we evaluate and compare the performance of the proposed GMERF method for point and MSE estimates with established SAE methods for binary target variables. In Section 1.5, we apply our proposed GMERF method to estimate area-level poverty rates and corresponding uncertainty estimates for the Mexican state of Tlaxcala. Section 1.6 presents a design-based simulation to evaluate the quality of the results obtained in Section 1.5, providing a comprehensive demonstration of the properties and advantages of GMERFs in the context of SAE. Finally, Section 1.7 concludes our study and motivates further research in the field of SAE.

## 1.2 Theory and method

In this section, we propose a novel, flexible, and data-driven approach that extends previous methods. Our method utilizes a random forest algorithm to estimate area-level proportions in the presence of unit-level survey and census data.

### 1.2.1 Generalized mixed effects random forests

Consider a finite population $U$ that is divided into $D$ areas. $U_i$ denotes the population of the $i$-th area with $i = 1, \ldots, D$ and $N_i$ denotes the population size of the $i$-th area. The overall population size is given by $N = \sum_{i=1}^{D} N_i$. The number of observations sampled in area $i$ is denoted by $n_i$ and the total sample size is given by $n = \sum_{i=1}^{D} n_i$. Within each sampled area, $j$ individual observations are obtained, where $j$ ranges from 1 to $n_i$. The binary response variable for area $i$ is represented by a vector of individual observations $y_i = [y_{i1}, \ldots, y_{in_i}]^\intercal$ with dimension $n_i \times 1$. Let $X_i = [x_{i1}, \ldots, x_{in_i}]^\intercal$ and $Z_i = [z_{i1}, \ldots, z_{in_i}]^\intercal$ represent the $n_i \times p$ matrix of covariates and the $n_i \times q$ matrix of domain-specific random effect specifiers, respectively. Here, $p$ denotes the number of covariates and $q$ denotes the dimension of the random effects. The $q \times 1$ vector of random effects for area $i$ is denoted by $\nu_i = [\nu_{i1}, \ldots, \nu_{iq}]^\intercal$ which we assume normally distributed with the variance-covariance matrix $H_i$ for random effects of each domain $i$. The notation used for all areas is as follows: $y = \text{col}_{1 \leq i \leq D}(y_i) = (y_i^\intercal, \ldots, y_D^\intercal)^\intercal$, $X = \text{col}_{1 \leq i \leq D}(X_i)$, $Z = \text{diag}_{1 \leq i \leq D}(Z_i)$, $H = \text{diag}_{1 \leq i \leq D}(H_i)$ and $\nu = \text{col}_{1 \leq i \leq D}(\nu_i)$. For example, in this notation, $\text{col}_{1 \leq i \leq D}(\nu_i)$ represents a column vector constructed by vertically stacking the elements $\nu_i$ for $i = 1, 2, \ldots, D$. Similarly, $\text{diag}_{1 \leq i \leq D}(Z)$ denotes a block diagonal matrix with blocks $Z_i$ along the diagonal and zeros elsewhere. The parameter of interest for each area, $\mu_i$, represents the true population-level mean of the binary response variable for area $i$ and is defined as:

$$\theta_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}.$$

The objective is to predict values for non-sampled observations using accessible supplementary covariates from census data across domains. To achieve this, we aim to establish a relationship $f()$ between the covariates and the target variable based on the survey data. We assume a non-informative survey design, meaning that the model for the sampled observations also holds for the non-sampled observations after conditioning on the covariates. The fixed part, denoted as $f()$, represents the random forest and expresses the conditional mean of the

linear predictor $\eta$ given the covariates $X$. The random part, represented by $Z\nu$, accounts for the dependencies introduced by random effects. $\mu$ denotes the vector consisting of the inverse logits of the elements of $\eta$. Following Pellagatti et al. (2021), the GMERF model is described in equation (1.1):

$$\begin{aligned} \eta &= f(X) + Z\nu, \\ \nu &\sim N(0, H), \\ \mu &= \frac{\exp(\eta)}{1 + \exp(\eta)}, \\ E(y|\nu) &= \mu. \end{aligned} \tag{1.1}$$

The model in equation (1.1) can be simplified to a generalized linear mixed model (GLMM) by setting $f(X) = X\beta$, where $\beta = [\beta_1, \ldots, \beta_p]^\intercal$ are the regression parameters. It is important to note that inference based on GLMMs poses computational challenges due to the involvement of high-dimensional integrals in the likelihood, which cannot be evaluated analytically (Stroup, 2012). To overcome this issue, a popular approximation method is the penalized quasi-likelihood (PQL) approach. The PQL approach constructs a linear approximation of the distribution for non-normal response variables and assumes that the linearized dependent variable is approximately normally distributed. This gives the integration a closed form, allowing the use of maximum likelihood estimation in the algorithm (Breslow and Clayton, 1993; Stroup, 2012). By employing the PQL approach to estimate a GLMM, we can derive a weighted MERF pseudo-model that uses a linearized target variable $y_L$. In order to linearize the binary response variable $y$, we apply a first-order Taylor series expansion, resulting in the transformed variable:

$$y_L = \ln\left(\frac{\mu}{1-\mu}\right) + (y - \mu)\left(\frac{1}{\mu(1-\mu)}\right). \tag{1.2}$$

The weighted MERF pseudo-model based on the MERF by Krennmair and Schmid (2022) is defined as follows:

$$y_L = f(X) + Z\nu + \epsilon, \tag{1.3}$$

with $\epsilon \sim N(0, W^{-1})$, $\epsilon = \text{col}_{1\leq i\leq D}(\epsilon_i)$, $\epsilon_i = [\epsilon_{i1}, \ldots, \epsilon_{in_i}]^\intercal$ is the $n_i \times 1$ vector of individual error terms and $W = \text{diag}_{1\leq i\leq D}(W_i)$ are the weights, where $W_i = \text{diag}_{1\leq j\leq n_i}(w_{ij})$ and $w_{ij} = \mu_{ij}(1-\mu_{ij})$. We assume that $\nu_i$ and $\epsilon_i$ are independent normally distributed, and the between-area observations are also independent. The covariance matrix of the linearized variable $y_L$ can be expressed as $Cov(y_L) = V = \text{diag}_{1\leq i\leq D}(V_i)$, with $V_i = Z_i H_i Z_i^\intercal + W_i^{-1}$. Additionally, we assume that correlations in the model (1.3) arise solely from the between-domain variation, i.e., $W_i^{-1}$ is diagonal. Note that in the special case of binary data, the variance-covariance matrix for the individual errors is equivalent to the inverse of the weights. These linearization weights are also used within the random forest algorithm for sampling training observations. They represent the influence of each observation on the model's fit and help prioritize those with higher variability during the training of the random forest. Sampling with probability proportional to the weight ensures that the bootstrap samples emphasize observations with

greater potential to improve the model's predictive performance. Unlike traditional survey sampling, where weights represent inverse first-order selection probabilities, the weights have a model-based interpretation. They account for the variability of the response variable and guide the random forest algorithm to focus on observations that contribute more significantly to estimating the binary outcome.

The proposed GMERF-algorithm for fitting model (1.1) is a doubly iterative process with micro iterations within macro iterations. In each macro iteration, the linearized response variable and weights are updated, while the micro iterations use an approach similar to the expectation-maximization (EM)-algorithm (Moon, 1996) for parameter estimates. The updated linearized response variable and weights values serve as response variable and weights, respectively:

1. Set $B = 0$ and assign initial estimates of the mean values $\hat{\mu}^{(B)}$ ($\hat{\mu}^{(B)} = 0.75$ for $y = 1$ and $\hat{\mu}^{(B)} = 0.25$ for $y = 0$). Calculate the linearized response $y_L^{(B)}$ by using equation (1.2).

2. Fit a weighted MERF pseudo model (1.3) using the linearized response $y_L^{(B)}$ and the weights $W^{(B)}$.

   (a) Initialize $b = 0$ and set random components $\hat{\nu}_{(0)}$ to zero.

   (b) Set $b = b + 1$. Update $y^*_{L(b)}$, $\hat{f}(X)_{(b)}$ and $\hat{\nu}_{(b)}$:

      i. $y^*_{L(b)} = y_L^{(B)} - Z\hat{\nu}_{(b-1)}$ (decorrelate the dependent variable).

      ii. Estimate $\hat{f}()_{(b)}$ using a random forest with dependent variable $y^*_{L(b)}$, covariates $X$ and weights $W^{(B)}$.

      iii. Get the Out-of-Bag-predictions (OOB-predictions) from the random forest $\hat{f}(X)^{OOB}_{(b)}$.

      iv. Estimate a linear mixed model (using maximum likelihood) with weights and restricted regression coefficient of 1 for $\hat{f}(X)^{OOB}_{(b)}$:

$$y_L^{(B)} = \hat{f}(X)^{OOB}_{(b)} + Z\hat{\nu}_{(b)} + \epsilon.$$

      v. Extract the variance components and estimate the random effects by:

$$\hat{\nu}_{(b)} = \hat{H}_{(b)} Z^\intercal \hat{V}^{-1}_{(b)} (y_L^{(B)} - \hat{f}(X)^{OOB}_{(b)}).$$

   (c) Repeat step (b) until convergence is reached in terms of generalized log-likelihood (GLL) value:

$$\begin{aligned} GLL(f, v_i|y) = -\sum_{i=1}^{D} [&(y_L - \hat{f}(X_i) - Z_i\hat{\nu}_i)^\intercal \hat{W}_i \\ &\times (y_L - \hat{f}(X_i) - Z_i\hat{\nu}_i) + \hat{\nu}_i^\intercal \hat{H}_i^{-1} \hat{\nu}_i \\ &+ \log|\hat{H}_i| + \log|\hat{W}_i^{-1}|]. \end{aligned}$$

3. Set $B = B + 1$: Update $\hat{\eta}$, $\hat{\mu}$, $\hat{y}_L$ and $w$:

$$\hat{\eta}^{(B)} = \hat{f}(X)^{OOB} + Z\hat{\nu},$$
$$\hat{\mu}^{(B)} = \frac{\exp(\hat{\eta}^{(B)})}{1 + \exp(\hat{\eta}^{(B)})},$$
$$y_L^{(B)} = \ln\left(\frac{\hat{\mu}^{(B)}}{1 - \hat{\mu}^{(B)}}\right) + (y - \hat{\mu}^{(B)})\left(\frac{1}{\hat{\mu}^{(B)}(1 - \hat{\mu}^{(B)})}\right),$$
$$W^{(B)} = \text{diag}_{1 \leq i \leq D}(W_i^{(B)}), \text{ where } W_i^{(B)} = \text{diag}_{1 \leq j \leq n_i}(w_{ij}^{(B)}) \text{ and } w_{ij}^{(B)} = \mu_{ij}^{(B)}(1 - \mu_{ij}^{(B)}),$$

where $\hat{f}(X)^{OOB}$ and $\hat{\nu}$ equal their estimated values at the micro-level convergence of the previous macro iteration.

4. Repeat steps 2 and 3 until $\hat{\eta}$ converges.

For given $V$ and $f = X\beta$, the maximization of the GLL-criterion is equivalent to the solution of the mixed model equations (Wu and Zhang, 2006). For given variance components, this leads to an estimator of the best linear unbiased predictor (BLUP) for the GMERF (González-Manteiga et al., 2007):

$$\hat{\nu} = HZ^{\intercal}V^{-1}(y_L - \hat{f}(X)). \tag{1.4}$$

### 1.2.2 Small area proportions

Assuming the same simplifications proposed by Battese et al. (1988) throughout the paper, i.e., $q = 1$, where $Z$ is a $n_i \times D$ design-matrix of area-intercept indicators, $\nu = [\nu_1, \ldots, \nu_D]^{\intercal}$ is a $D \times 1$ vector of random effects, and the variance-covariance matrix of random effect simplifies to $H_i = \sigma_\nu^2$. Since $\hat{\nu}_i$ is the BLUP, the proposed estimator for the area-level proportion is given by:

$$\hat{\eta}_{ij} = \hat{f}(x_{ij}) + \hat{\nu}_i,$$
$$\hat{\mu}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\frac{\exp(\hat{\eta}_{ij})}{1 + \exp(\hat{\eta}_{ij})} \text{ for } i = 1, \ldots, D. \tag{1.5}$$

For non-sampled areas, the proposed estimator for the area-level proportion simplifies to the fixed component obtained from the random forest:

$$\hat{\eta}_{ij} = \hat{f}(x_{ij}),$$
$$\hat{\mu}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\frac{\exp\hat{\eta}_{ij})}{1 + \exp(\hat{\eta}_{ij})}.$$

Note that $x_{ij}$ represents information on $p$ covariates and is available for all $N$ units in the population $U$ from census micro-data. Although this may be a restrictive assumptions in some applications, the covariates $x_{ij}$ cannot be directly substituted by aggregated covariates - unlike

in linear models - to construct the estimator for the proportion for area $i$. This is because

$$\bar{\hat{f}}(X_i) = \frac{1}{N_i}\sum_{j=1}^{N_i}\hat{f}(x_{ij}) \neq \hat{f}\big(\frac{1}{N_i}\sum_{j=1}^{N_i}x_{ij}\big) = \hat{f}(\overline{X}_i).$$

Unlike existing SAE approaches for binary target variables, our proposed estimator circumvents issues related to model selection. This advantage stems from the utilization of the random forest method, which implicitly incorporates optimized model selection, encompassing higher-order effects and non-linear interactions. Moreover, our estimator is adept at handling high-dimensional covariate data, even when the number of covariates surpasses the sample size (Hastie et al., 2009). The predictive performance of our method is influenced by two crucial tuning parameters: the number of split-candidates at each node, which regulates the degree of decorrelation, and the number of trees. These parameters are selected through repeated cross-validation.

The GMERF framework adopts a plug-in prediction approach, which sets it apart from methods that directly compute the conditional expectation, such as the EBP. Instead of deriving predictions through explicit integration over random effects, GMERF iteratively estimates parameters in the fixed and random components. This process substitutes true parameters with their estimates, allowing the method to approximate area-level proportions efficiently. The decision to use a plug-in predictor ensures its usability in data-rich, computationally intensive contexts. Although this approach may not achieve the theoretical optimality of EBPs, it represents a practical and effective solution for addressing the challenges of SAE with binary outcomes.

In the subsequent discussion, we aim to explore the data situations in which our GMERF approach (for binary variables) may exhibit great performance and the potential to outperform the MERF (for continuous variables) when the interest is in estimating area-level proportions. This investigation will be carried out through application and design-based simulations in Section 1.5 and Section 1.6, focusing on comparing the performance of our proposed approach with established SAE methods that utilize continuous information, which inherently provides more information for the target variable. The PQL approach used in our algorithm works particularly well when the random effects are normally distributed (McCulloch, 1997). However, Krennmair and Schmid (2022) demonstrated the robustness of the MERF (whose properties we incorporate into our method) against misspecification of distributional assumptions in mixed models. Given that the GMERF relies on an approximation method, it is advisable to follow a rule of thumb similar to the central limit theorem when approximating a binary variable using the normal distribution. Finally, we can draw upon an argument from the classical linear model (without random effects). In certain cases, a binary indicator is estimated using a linear probability model (LPM) that does not rely on a link function. This procedure is viable because the LPM often exhibits comparable performance to the traditional logit model, particularly when the probabilities fall within a specific range where they exhibit almost linear behavior with respect to the log-odds function. Generally, the LPM can be employed when the estimated probabilities range lies between $0.2$ and $0.8$ (Long, 1997). Hence, we suspect that our approximation method exhibits satisfactory performance when the area-level probabilities

predominantly fall within this range.

## 1.3 Uncertainty estimation

In the context of small area estimators, having an accuracy measure, typically expressed as MSE, is crucial. However, computing the analytical form of the MSE is not feasible even for relatively simple models like GLMMs. To address this, González-Manteiga et al. (2007) propose a rough approximation by linearizing the model and applying the Prasad-Rao approximation (Prasad and Rao, 1990) for linear mixed models. Alternatively, bootstrap schemes provide a straightforward approach. In this paper, we propose a parametric bootstrap to estimate the MSE of the small area estimator introduced in equation (1.5). Our bootstrap scheme is built upon the method proposed by González-Manteiga et al. (2007) for small area estimators based on a logit mixed model. This approach focuses on capturing the dependence structure of the data and the uncertainty arising from the model estimation. The bootstrap procedure consists of the following steps:

1. For $b = 1, \ldots, B$:
   (a) Generate bootstrap random effects for the $D$ areas as $\nu_i^{(b)} \sim N(0, \hat{\sigma}_\nu^2)$ for $i = 1, \ldots, D$.
   (b) Generate a bootstrap population of independent Bernoulli realizations made up of $D$ areas of sizes $N_i$, and with bootstrap Bernoulli realization $y_{ij}^{(b)}$ in area $i$ taking the value 1 with probability:

$$\mu_{ij}^{(b)} = \frac{\exp\big(\hat{f}(x_{ij}) + \nu_i^{(b)}\big)}{1 + \exp\big(\hat{f}(x_{ij}) + \nu_i^{(b)}\big)}.$$

   (c) Calculate the true bootstrap population area probabilities $\mu_i^{(b)}$ as $\frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}^{(b)}$ for all $i = 1, \ldots, D$.
   (d) 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 $\hat{\nu}^{(b)}()$.
   (e) Calculate area-level probabilities by:

$$\hat{\mu}_i^{(b)} = \frac{1}{N_i}\sum_{j=1}^{N_i} \frac{\exp\big(\hat{f}^{(b)}(x_{ij}) + \hat{\nu}_i^{(b)}\big)}{1 + \exp\big(\hat{f}^{(b)}(x_{ij}) + \hat{\nu}_i^{(b)}\big)}.$$

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

$$\widehat{MSE}_i^{param} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\mu}_i^{(b)} - \mu_i^{(b)})^2.$$

## 1.4 Model-based simulation study

This section marks the initial phase of our empirical evaluation of the proposed method. To assess its performance, we employ a model-based simulation that compares point estimates for

area-level proportions derived from the GMERF model in equation (1.1) with those obtained from several competing models. Specifically, we investigate the performance of GMERFs in comparison to established SAE methods for binary outcomes, such as the M-quantile (MQ) approach (Chambers et al., 2016) and the conditional expectation predictor (CEP). The CEP, in this context, is the plug-in predictor based on a logistic mixed model:

$$\hat{\mu}_i^{CEP} = \frac{1}{N_i}\left(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{\mu}_{ij}\right),$$

with $\hat{\mu}_{ij} = \exp(x_{ij}^\intercal \hat{\beta} + \hat{\nu}_i)(1 + \exp(x_{ij}^\intercal \hat{\beta} + \hat{\nu}_i))^{-1}$ (González-Manteiga et al., 2007). By contrasting the performance of these linear competitors with our more flexible approach, which incorporates semi-parametric and non-linear modeling, we aim to showcase the advantages of our methodology. The primary objective is to demonstrate that our proposed method, in terms of point and uncertainty estimates, performs comparably well to traditional SAE methods while exhibiting comparative strengths in terms of capturing non-linear relations in the data.

The simulation setting is defined by a finite population $U$ of size $N = 50000$, comprising $D = 50$ disjoint areas $U_1, ..., U_D$ of equal size $N_i = 1000$. To generate samples, we employ stratified random sampling, treating the $50$ small areas as strata. This results in a sample size of $n = \sum_{i=1}^{D} n_i = 687$. The sample sizes for each area range from 1 to 28 sampled units, with a median of 13 and a mean of 14. These sample sizes align with the area-level sample sizes observed in the application described in Section 1.5.

We contemplate four scenarios: *Normal-Small*, *Interaction-Small*, *Normal-Large*, and *Interaction-Large*. Each scenario is independently repeated $M = 500$ times. By comparing the estimates obtained from different models in these scenarios, we can assess their performance when faced with unknown non-linear interactions between covariates and varying levels of variation in the random effect. The *Normal* scenarios serve as a reference for the CEP and MQ models. Since the model assumption of linearity is fulfilled, we aim to show that GMERFs perform comparably well to linear competitors in the reference scenario. The *Interaction* scenarios differ from the *Normal* ones as they involve a more complex model that incorporates quadratic terms and interactions on the linear predictor scale. These scenarios demonstrate the advantages of semi-parametric and non-linear modeling methods that protect against model-failure. Within each of these two scenarios, we explore two specifications for the random effects: small and large. These specifications account for different levels of magnitude in the between-group variability, thereby providing insights into the performance of the models under varying levels of variability. The indication of a small or large random effect is always to be put in relation to the given effect strength of the fixed effects. To illustrate this more precisely, we calculate the variance partition coefficient (VPC) (Goldstein et al., 2002). The VPC serves as a measure of intraclass correlation. In the context of binary outcomes, it indicates the explanatory variance share on the target variable that is to be assigned to the classification by the group variable. The VPC can be calculated as follows:

$$VPC = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_{latent}^2},$$

where $\sigma_{latent}^2$ denotes the residual variance that cannot be explained by either the fixed effects

or the group-specific random intercepts. This latent variance can be estimated by the variance of the logistic distribution, which is $\sigma^2_{latent} = \frac{\pi^2}{3}$, with $\pi \approx 3.14159$, as demonstrated in the fixed effects logistic regression model (Browne et al., 2005; Fontana et al., 2021). In this study, we use the terms small or large to describe scenarios with a VPC of $0.03$ or $0.23$, respectively. In all scenarios, the binary response variable $y$ follows a Bernoulli distribution with success probability $\mu$, where $\mu = \exp(\eta)(1 + \exp(\eta))^{-1}$. The covariates $x_1$ and $x_2$ are independently drawn from $N(0, 2^2)$ and $N(0, 3^2)$, respectively. The linear predictor $\eta$ and the random effects $\nu$ for each scenario are detailed in Table 1.1.

Table 1.1: Model-based simulation scenarios

| Scenario | Linear predictor | $\nu$ |
|---|---|---|
| Normal-Small | $\eta = 0.5 - 0.8x_1 - 0.6x_2 + \nu$ | $N(0, 0.1)$ |
| Interaction-Small | $\eta = 1 - x_1x_2 - 0.6x_1^2 + \nu$ | $N(0, 0.1)$ |
| Normal-Large | $\eta = 0.5 - 0.1x_1 - 0.2x_2 + \nu$ | $N(0, 1)$ |
| Interaction-Large | $\eta = 1 - x_1x_2 - 0.6x_1^2 + \nu$ | $N(0, 1)$ |

We assess the quality of point estimates for area-level proportions using two metrics: relative bias (RB) and relative root mean squared error (RRMSE). To evaluate the accuracy of the MSE estimates, we consider the relative bias of the root mean squared error (RB-RMSE) and the relative root mean squared error of the RMSE. These quantities are defined as

$$
\begin{aligned}
RB_i &= \frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{\mu}_i^{(m)} - \mu_i^{(m)}}{\mu_i^{(m)}}\right) \\
RRMSE_i &= \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\mu}_i^{(m)} - \mu_i^{(m)}\right)^2}}{\frac{1}{M}\sum_{m=1}^{M}\mu_i^{(m)}} \\
RB\text{-}RMSE_i &= \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M} MSE_{est_i}^{(m)}} - RMSE_{emp_i}}{RMSE_{emp_i}} \\
RRMSE\text{-}RMSE_i &= \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\sqrt{MSE_{est_i}^{(m)}} - RMSE_{emp_i}\right)^2}}{RMSE_{emp_i}},
\end{aligned}
$$

where $\hat{\mu}_i^{(m)}$ represents the estimated proportion in area $i$ obtained from any of the methods mentioned above and $\mu_i^{(m)}$ denotes the true proportion for area $i$ in simulation round $m$. The estimation of $MSE_{est_i}^{(m)}$ is performed using the proposed bootstrap method described in Section 1.3. Furthermore, $RMSE_{emp_i}$ is calculated as the empirical root mean squared error over $M$ replications, given by $\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\mu}_i^{(m)} - \mu_i^{(m)})^2}$. Since MSE estimators are commonly used to construct confidence intervals (CIs) for small area quantities of interest, we also assess the coverage properties of these intervals. Following Chandra et al. (2018), we examine the 95% confidence intervals and compute the percentage coverage as:

$$
CR_i = 100 \cdot \left[\frac{1}{M}\sum_{m=1}^{M}\mathbb{1}\left(\left|\hat{\mu}_i^{(m)} - \mu_i^{(m)}\right| \leq 1.96\sqrt{MSE_{est_i}^{(m)}}\right)\right].
$$

To implement the model-based simulation, we use R (R Core Team, 2024). The CEP estimates are obtained using the **lme4** package (Bates et al., 2015), while the **BinaryMQ** package (Chambers et al., 2016) is used to compute the MQ estimates. For the proposed GMERF approach, we employ the **ranger** (Wright and Ziegler, 2017) and **lme4** (Bates et al., 2015) packages. To monitor the convergence of the algorithm, we use a precision of $1e^{-5}$ in the relative difference of the GLL-criterion and a precision of $0.01$ in the relative change of $\hat{\eta}$.

Figure 1.1 displays the empirical RMSE of each method across the four scenarios. As anticipated, in the *Normal-Small* scenario, the GMERF method does not surpass the CEP and MQ estimators, but instead performs comparably. A similar pattern is observed in the *Normal-Large* scenario, with all three estimators exhibiting a higher overall RMSE. The competitors conform to the fixed effects of the data-generating process and perform accordingly. For complex scenarios, namely *Interaction-Small* and *Interaction-Large*, point estimates from the proposed GMERF method outperform the competitors. In the *Interaction-Small* scenario, the CEP has lower RMSE values than the MQ estimator. However, in the *Interaction-Large* scenario, the MQ estimator outperforms the CEP in terms of lower RMSE. The flexible GMERF approach automatically identifies interactions and non-linear relationships, such as quadratic terms, leading to an advantage in terms of RMSE. Overall, the results in Figure 1.1 indicate that the GMERF performs comparably well in linear scenarios and outperforms traditional SAE-models in the presence of unknown non-linear relationships. Table 1.2 reports the corresponding values of RB and RRMSE for each discussed point estimate. The RB and RRMSE of the GMERF method exhibit a competitive low level across all scenarios. Specifically, in the *Interaction-Large* scenario, a well-known observation regarding the statistical properties of random forests becomes apparent: the RB is slightly higher compared to the CEP, yet this increased RB is compensated by a lower RRMSE for point estimates. This behavior reflects the bias-variance trade-off inherent in random forests (Hastie et al., 2009). By averaging predictions across an ensemble of trees, random forests significantly reduce variance, resulting in more stable predictions. However, this averaging also introduces bias, as predictions are regularized toward the mean.

Table 1.2: Mean and median of RB and RRMSE over areas for point estimates.

| | *Normal-Small* | | *Interaction-Small* | | *Normal-Large* | | *Interaction-Large* | |
|---|---|---|---|---|---|---|---|---|
| | Median | Mean | Median | Mean | Median | Mean | Median | Mean |
| RB[%] | | | | | | | | |
| CEP | 0.5737 | 0.6575 | 1.3044 | 1.1705 | 5.0801 | 5.7415 | 1.2549 | 1.2620 |
| GMERF | 0.2685 | 0.2939 | 0.0683 | 0.0777 | 5.1449 | 5.6290 | 1.5751 | 1.6645 |
| MQ | 2.1991 | 2.2473 | 2.4214 | 2.3913 | 7.7560 | 7.9638 | 4.5263 | 4.6023 |
| RRMSE[%] | | | | | | | | |
| CEP | 10.3583 | 10.4083 | 15.6118 | 15.7275 | 27.3135 | 29.9692 | 23.7271 | 23.7545 |
| GMERF | 10.0811 | 10.1047 | 10.4459 | 10.5813 | 31.1783 | 33.1685 | 18.2625 | 18.3164 |
| MQ | 10.0061 | 10.9667 | 19.4710 | 21.4834 | 33.4434 | 35.8279 | 21.9926 | 24.6999 |

We now assess the performance of the MSE estimator implemented with the parametric bootstrap method presented in Section 1.3 with $B = 200$ bootstrap replications. In Table

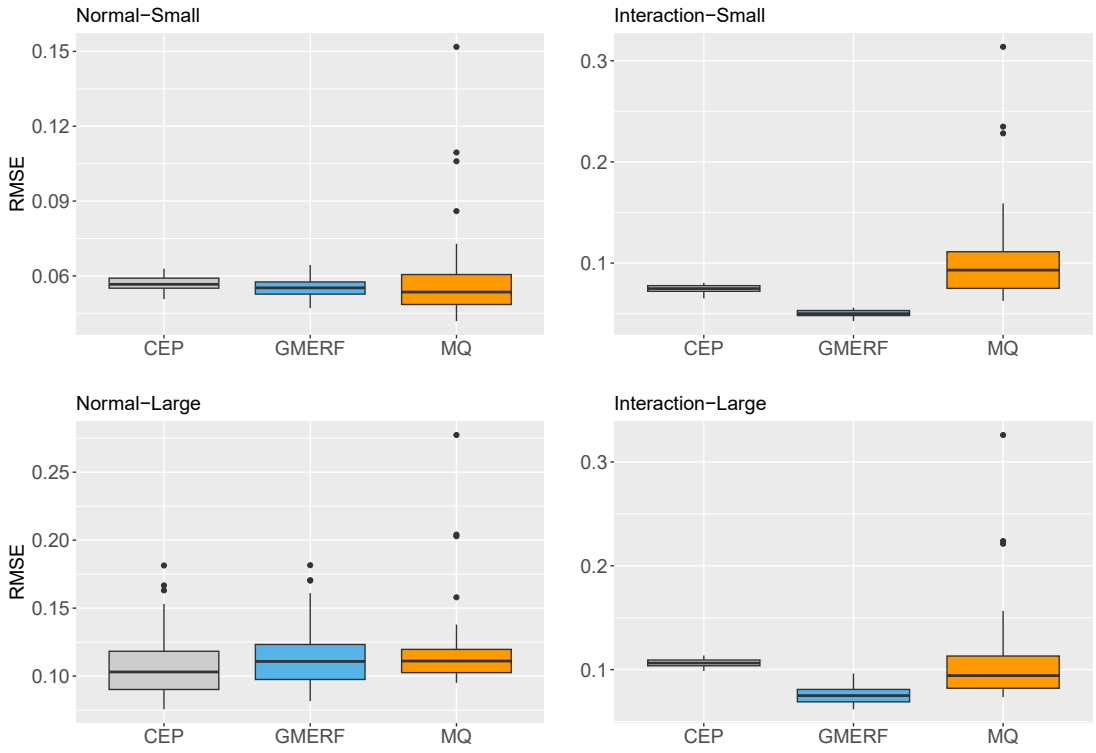



Figure 1.1: Empirical RMSE comparison of point estimates for area-level proportions under four scenarios.

1.3, we observe the RB-RMSE of the proposed parametric bootstrap procedure across the four scenarios. In particular, the proposed MSE estimator demonstrates reasonably low relative bias in terms of mean and median values over areas under all four scenarios. Although we cannot directly infer the area-wise tracking properties of the estimated RMSE against the empirical RMSE from the results of Table 1.3, Figure 1.2 provides additional insight into the quality of our proposed parametric MSE-bootstrap estimator. Based on the tracking properties in all four scenarios, we conclude that using the parametric bootstrap for estimating the MSE appears to have appealing properties regarding bias and stability. Furthermore, we present the percentage coverage rates of the proposed parametric bootstrap procedure across the four scenarios in Table 1.3. In terms of the coverage characteristics of the nominal 95% CIs, the bootstrap MSE estimator slightly underestimates the coverage rates. Overall, the proposed MSE estimator demonstrates satisfactory performance and effectively estimates the true MSE of the GMERF.

Table 1.3: Performance of bootstrap MSE estimators in model-based simulation: mean and median of RB-RMSE, RRMSE-RMSE and coverage rate over areas

| | *Normal-Small* | | *Interaction-Small* | | *Normal-Large* | | *Interaction-Large* | |
|---|---|---|---|---|---|---|---|---|
| | Median | Mean | Median | Mean | Median | Mean | Median | Mean |
| RB-RMSE[%] | −9.17 | −9.23 | −2.78 | −2.28 | −4.91 | −5.20 | −3.11 | −3.08 |
| RRMSE-RMSE[%] | 25.36 | 25.72 | 24.84 | 25.40 | 8.23 | 8.75 | 24.73 | 24.98 |
| CR[%] | 89.54 | 89.40 | 91.43 | 91.44 | 93.70 | 93.65 | 89.16 | 89.14 |

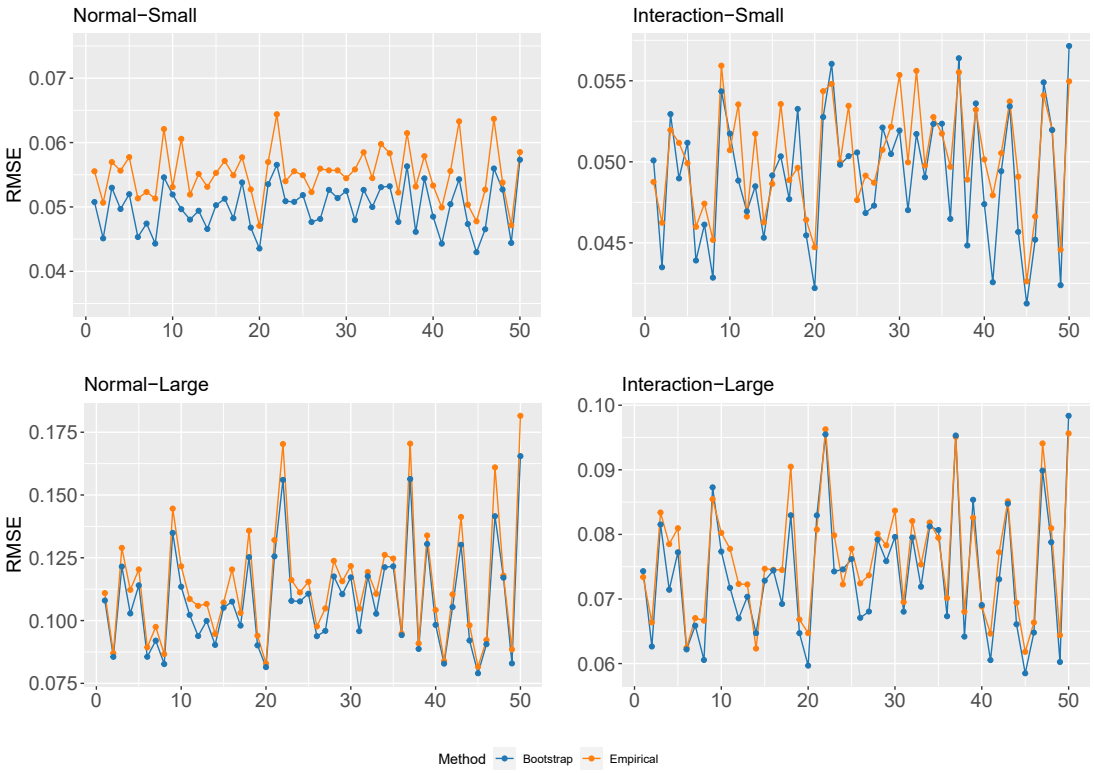



Figure 1.2: Empirical and bootstrapped area-level RMSEs for four scenarios

## 1.5 Application: Estimating poverty rates for Mexican municipalities

In this section, we investigate the performance implications of estimating the HCR using a binary variable versus a continuous income scale. Concretely, we compare the performance of the GMERF method with established methods for estimating poverty indicators in the context of SAE that use continuous household income as an input variable: the EBP with Log-Shift transformation (Rojas-Perilla et al., 2020) and the MERF for non-linear indicators (Krennmair, 2022). In addition, we incorporate the Fay-Herriot model with a logit transformation (Runge, 2023). The estimator, referred to as FH logit, operates on aggregated survey data (direct estimates) and census means. It models the data at the aggregated level and implicitly accounts for survey weights through the direct estimates. We apply these methods to Mexican income data to estimate area-specific HCRs and their associated uncertainties. Section 1.5.1 describes the data, while Section 1.5.2 presents the results.

### 1.5.1 Data description

Mexico has been plagued by poverty, affecting millions of children, men, women, the elderly, and the indigenous population in particular (Consejo Nacional de Evaluación de la Política de Desarrollo Social, 2020). The country faces various obstacles in its development, including crime, governmental opacity, a scarcity of skilled labor, and corruption. Nearly half of the population lives below the poverty line, making poverty reduction a crucial priority. Since 2008,

Mexico has implemented a multidimensional poverty measurement system, providing a more accurate assessment of social development policies at the federal, state, and municipal levels (Organisation for Economic Co-operation and Development, 2012). However, the country has achieved mixed results in terms of reducing poverty rates. The social rights dimension has seen progress in terms of basic service coverage, such as education, health, housing, and social security. Meanwhile, the economic well-being dimension, as captured by individuals' wages, has experienced fluctuations over time, driven by significant events such as the financial crisis of 2008-2009. Subsequently, there was a period of recovery observed between 2014 and 2018. Nevertheless, the onset of the COVID-19 pandemic led to a subsequent decrease in economic well-being indicators (Statista, 2021). The picture is heterogeneous across the federal states, and this poses differentiated challenges to policymakers. Efforts have been made to refine the conceptual and methodological framework for measuring poverty, understand the living conditions of the population in poverty, and precisely identify their geographical location. As part of the Mexican State's institutional efforts to overcome poverty, the National Institute of Statistics and Geography (INEGI) designated the National Survey of Household Income and Expenditure (ENIGH) and the Socioeconomic Conditions Module of the ENIGH (MCS-ENIGH) as information of national interest. These data sources provide the necessary sample and auxiliary data required for estimating HCRs using SAE techniques. The ENIGH from 2010 serves as the sample data, while the census of 2010 provides the auxiliary data for the analysis. This study examines the regional differences in HCRs within one of the 32 Mexican federal states, Tlaxcala. Tlaxcala consists of 60 municipalities and had a population of 1342977 in 2020. It is the fifth least populated federal state, covering an area of 3996.6 km$^2$ and having a population density of 336 inhabitants per km$^2$ (Instituto Nacional de Estadística y Geografía, 2021). Additional geographic information (including the names of the municipalities) is represented in Figure A.1 and Table A.1.

The target variable for the EBP Log-Shift and MERF models is the total household per capita income (*ictpc*), measured in pesos, which is available in the survey but not in the census. For the GMERF model, a binary target variable indicating whether a household is considered poor is created using the poverty line $t = 0.60 \times median(\textit{ictpc})$, where households with $\textit{ictpc}_{ij} \leq t$ are classified as poor. For the FH logit, the target variable is the direct estimator (Horvitz and Thompson, 1952) of the HCR, defined as

$$\widehat{\text{HCR}}_i = \frac{1}{\sum_{j=1}^{n_i} \omega_{ij}} \sum_{j=1}^{n_i} \omega_{ij} \mathbb{1}\big(\textit{ictpc}_{ij} \leq t\big),$$

where $\omega_{ij}$ represents the survey weight of a household $j$ in municipality $i$.

The census data for Tlaxcala from $2010$ contains information on $57751$ households and the survey data contains information on $1667$ households. Of the 60 municipalities, 52 are used in-sample and 8 are out-of-sample. Details on survey and census data properties are provided in Table 1.4. Regarding the survey, the maximum sample size of a municipality is $143$, the minimum is $2$ and the median is $21.50$ households per municipality. For the EBP Log-Shift and FH logit models, we follow the approaches of Rojas-Perilla et al. (2020) and Schmid et al. (2017), respectively, and use the Bayesian Information Criterion (BIC) to identify valid predictors for

the target variables. Table A.2 in the Appendix provides a summary of the variables used for both methods. The GMERF algorithm's convergence is monitored under a precision of $1e^{-5}$ in the relative difference of the GLL criterion and with a precision of $0.01$ in the relative change in $\hat{\eta}$. The default of 500 trees is retained, and based on 5-fold cross-validation on the original survey-sample, it is recommended to use 2 variables at each split for the GMERF. Figure A.2 in the Appendix presents partial dependence plots, providing motivation for exploring the use of the GMERF method in this study. These plots estimate the marginal effect of covariates on the target variable, helping to assess whether the relationship tends to be linear or more complex. Overall, Figure A.2 reveals both linear and non-linear relationships, underscoring the importance of employing a flexible approach like random forests for this application.

Table 1.4: Summary statistics on in- and out-of-sample areas:
area-specific sample size of census and survey data.

| | Total | | In-sample | | Out-of-sample | |
|---|---|---|---|---|---|---|
| | 60 | | 52 | | 8 | |
| | Min. | Q1 | Median | Mean | Q3 | Max. |
| Survey area sizes | 2.00 | 11.00 | 21.50 | 32.06 | 40.25 | 143.00 |
| Census area sizes | 611.00 | 772.20 | 912.00 | 962.59 | 1,100.20 | 3,310.00 |

In Section 1.5.2 and the subsequent design-based simulation in Section 1.6, we present an illustrative and realistic example of the estimation of area-level HCRs using data from Tlaxcala. Our choice of this challenging example is deliberate, as it serves to demonstrate the validity of our proposed approaches for point and uncertainty estimates. Specifically, we aim to show that our method provides a viable alternative to existing SAE methods, while requiring less information about the input variables.

### 1.5.2 Results and discussion

Figure 1.3 displays the results of direct estimates, as well as those from the model-based estimation methods (EBP Log-Shift, FH logit, MERF and GMERF). Direct estimation of the HCRs is possible for 52 out of 60 domains. Clearly, the model-based estimates expand our understanding of regional disparities in HCRs beyond the sampled areas. All four model-based methods reveal distinct regional differences among municipalities. The state capital, Tlaxcala, located in the central Mexican highlands, has the lowest poverty rate. Income from tourism, given the state's rich history, could be one explanation for this phenomenon. Municipalities with slightly higher HCRs tend to be those with small commercial activities, such as Chiautempan, Humantla, San Pablo del Monte, and Zacatelco. Those with similarly high HCRs tend to be those with light industry, such as Atlangatepec, Calpulalpan, Ixtacuixtla de Mariano Matamoros, Nanacamilpa de Mariano Arista, Xicohtzinco, and Santa Isabel Xiloxoxtla, producing clothing, foam and plastic products, paper products, publishing, textiles, and automobiles. Municipalities in the north and east of the state tend to have the highest poverty rates, likely due to their rural nature and reliance on agriculture, livestock, forestry, and fishing (Instituto Nacional de Estadística y Geografía, 2017). Overall, the four model-based estimates

consistently depict the geographic distribution of HCRs, regardless of whether they are based on a binary household-level variable (GMERF and FH logit) or a continuous household-level variable (MERF and EBP Log-Shift).

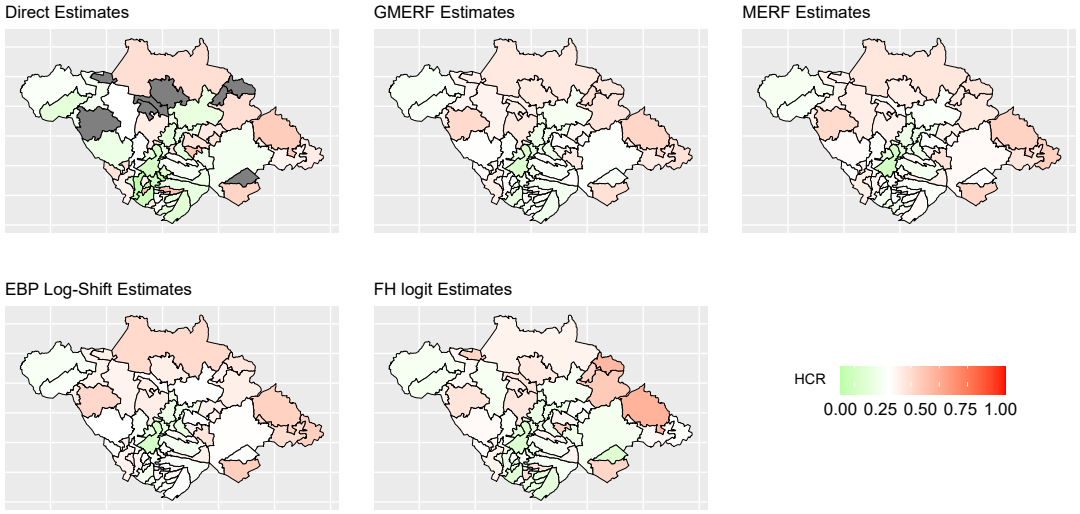



Figure 1.3: Estimated HCRs for the state Tlaxcala based on five different estimation methods.

In addition to mapping the empirical results of domain proportions, our focus is on quality criteria, particularly the coefficients of variation (CV). We use the calibrated bootstrap method provided in the R package **emdi** (Kreutzmann et al., 2019) to obtain estimates of variances for the direct estimates. To estimate the MSE for the EBP Log-Shift, FH logit and MERF, we employ the wild-bootstrap, parametric bootstrap and non-parametric bootstrap, respectively, as proposed by Rojas-Perilla et al. (2020), Runge (2023) and Krennmair and Schmid (2022). For the GMERF, we rely on the parametric bootstrap from Section 1.3, with $B = 200$ bootstrap replications. To evaluate the stability of our bootstrap approach for GMERF, we further examine the effect of the number of bootstrap replications on the estimated MSE in our application. Figure A.3 in the Appendix illustrates this relationship for eight municipalities in Tlaxcala. The results indicate that the estimated MSE stabilizes at approximately 200 replications, as further increasing $B$ does not lead to substantial variations in the MSE estimates. The corresponding CVs for in- and out-of-sample domains are reported in Figure 1.4. We note an improvement in the in-sample CVs for the EBP Log-Shift, FH logit, GMERF, and MERF compared to the CVs for the direct estimates. The median CVs for the EBP Log-Shift and GMERF are lower than those for the MERF and FH logit. In terms of CVs for out-of-sample areas, we found that our proposed GMERF approach has an advantage. However, upon analyzing individual CV values, it is unclear whether the improved performance of GMERFs is due to superior point estimates for domain-level proportions or its relatively lower MSE-estimates.

Thus, Figure 1.5 compares point estimates of direct estimates to model-based estimates for both in-sample and out-of-sample domains. Notably, there are no discernible systematic differences between the estimates from the EBP Log-Shift, GMERF, and the MERF. The in-sample areas in Figure 1.5 are sorted by decreasing sample sizes. In comparison to the direct and FH logit estimates, the predicted HCRs of EBP Log-Shift, GMERF, and MERF exhibit

less variation due to the impact of shrinkage.

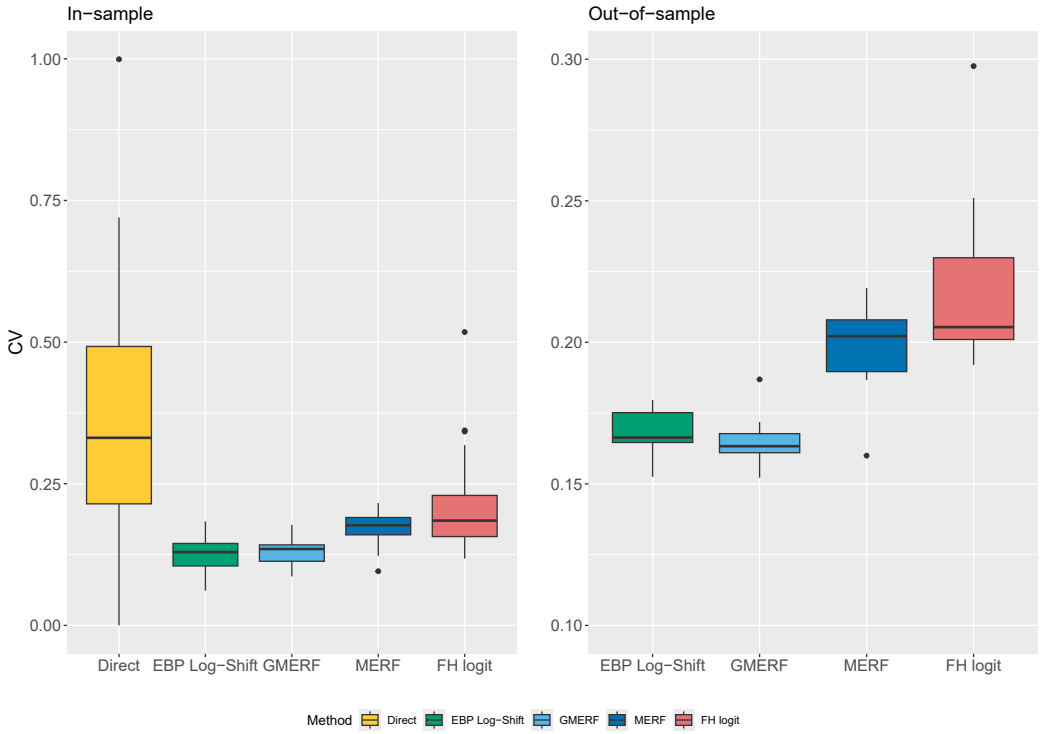



Figure 1.4: Domain-specific CVs for HCRs for in- and out-of-sample domains.


Figure 1.5: Detailed comparison of point estimates for the area-level HCRs. The dotted line separates sampled from non-sampled areas. In-sample areas are sorted by decreasing sample size.

Finally, we conduct an informal evaluation following the approach of Tzavidis et al. (2018) to compare model-based and design-based point estimates for aggregated geographical levels, providing an indication of the quality of the model-based estimates. Tlaxcala consists of 60 municipalities and 15 districts. The sample sizes range from 11 to 169 households per district, with a mean of 111 and a median of 109. We compare model-based estimates with design-based estimates for 14 districts for which the CVs of the design-based estimates are below 30%. Figure 1.6 displays point estimates for district-level HCRs obtained using the direct estimator, EBP Log-Shift, FH logit, MERF, and GMERF. Direct estimates are based on district-specific samples, while model-based estimates are aggregated from the corresponding municipality-level estimates using weights defined by $N_i/N$, where $N_i$ denotes the municipality population size. The districts are ordered by decreasing CVs of the direct estimates from left to right. We observe that, for districts where the direct estimates are less reliable (left-hand side of the plot), the model-based estimates diverge more from the direct estimates. In contrast, for districts where the design-based estimates are more reliable (right-hand side of the plot), the GMERF estimates tend to be closer to the direct estimates. The correlation coefficient between the direct estimator and the GMERF (0.87) on district-level is higher than the correlation between the direct estimator and the EBP Log-Shift (0.78) and MERF (0.76). Notably, the highest correlation is observed between the FH logit and the direct estimator (0.90), which is expected given that the FH logit models the direct estimates as target variable. A design-based simulation will further validate our results and facilitate a more detailed discussion of our method.

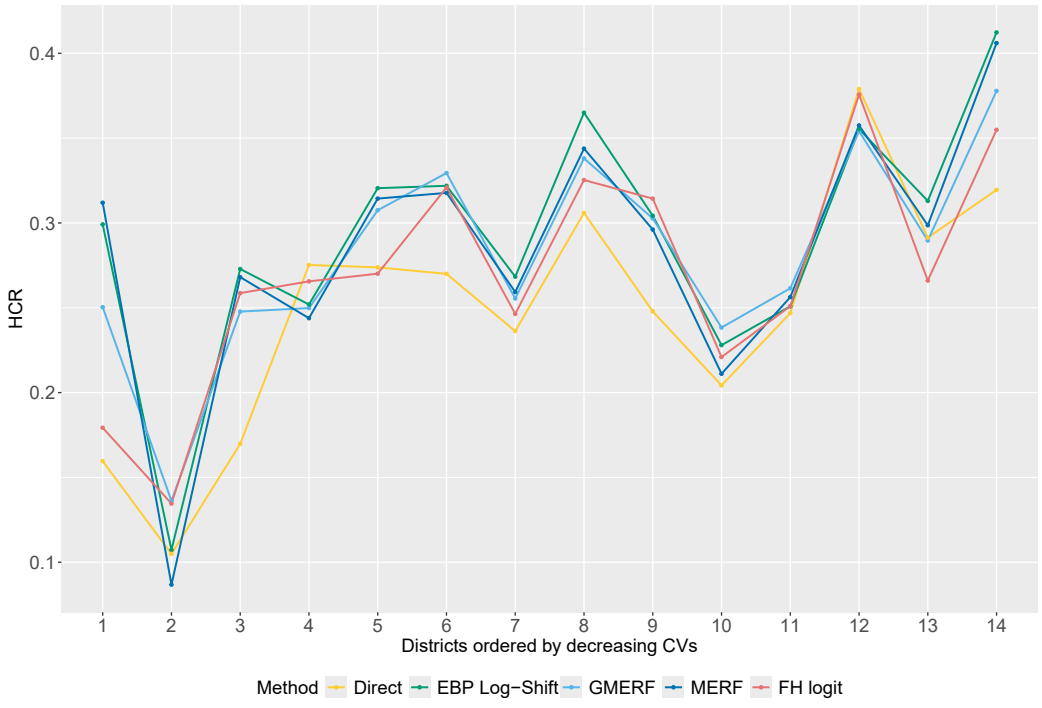



Figure 1.6: Estimates for HCRs at district-level.

## 1.6 Design-based simulation study

Regarding the design-based simulation, we are fortunate to have a variable in the survey and census datasets that is highly correlated with *ictpc* in our application: the variable *inglabpc*, which measures earned per capita income from work. Although this variable only covers one aspect of total household income and deviates from our desired income definition, it is effective in evaluating our method within a design-based simulation. By directly comparing the performance of the proposed GMERF approach to existing SAE methods for the estimation of area-level HCRs based on empirical data, the design-based simulation evaluates the results from the application.

We focus on area-level HCRs in the Mexican state of Tlaxcala and use the same data as in Section 1.5.2. We draw 500 independent pseudo-samples from the fixed population, maintaining the original survey's municipality sample sizes. This results in $500$ equally structured samples, each with an overall size of $1667$ households, as in the ENIGH 2010 survey. The true values are defined as the area-level HCRs from the fixed population/ census. We employ the same methods as described in the application in Section 1.5.2 and use consistent working models for EBP Log-Shift and FH logit, assuming they remain fixed throughout the design-based simulation. Additionally, for GMERF and MERF, we use the same tuning parameters discussed in Section 1.5.2. These include, for example, the number of trees in the random forest, the number of variables considered at each node for splitting, and the convergence criteria used in the iterative fitting process. As an additional comparative estimator, we implement the CEP (as in the model-based simulation from Section 1.4) to investigate whether the GMERF also has an advantage over the logistic mixed model when using real data.

To start with, we examine the effectiveness of our method in correctly classifying households into the 'true' category (poor/rich) using the receiver operating characteristic (ROC) and calibration curves presented in Figure 1.7. The ROC curve illustrates the relationship between the true positive rate (sensitivity) and the false positive rate ($1-$specificity), while the area under the curve (AUC) indicates the model's ability to differentiate between the two classes. A value closer to 1 indicates better class separation and prediction performance. In our case, we achieved an AUC of 0.8858, suggesting high classification and prediction accuracy. Additionally, the calibration curve on the right-hand side of Figure 1.7 displays the model's ability to accurately predict probabilities, by comparing the mean of predicted probabilities against the corresponding true probabilities at different quantiles. A diagonal line indicates a perfect match between the predicted and true probabilities. Our method demonstrates excellent performance in predicting probabilities within the $0.25$ to $0.75$ range, consistent with our theoretical considerations in Section 1.2.

We begin by discussing the performance of our method in terms of point estimation, comparing the CEP, EBP Log-Shift, FH logit, GMERF, and MERF. Figure 1.8 displays the average RMSE of the area-level HCRs for Tlaxcala, both overall and split by the 52 in-sample and 8 out-of-sample areas. The GMERF method produces the lowest median RMSEs. However, for out-of-sample areas, there is only a small difference in the RMSEs of GMERF and MERF. Given the high share of sampled municipalities in Tlaxcala, the in-sample areas' performance is a critical indicator of each method's quality and stability. In this case, GMERF and MERF

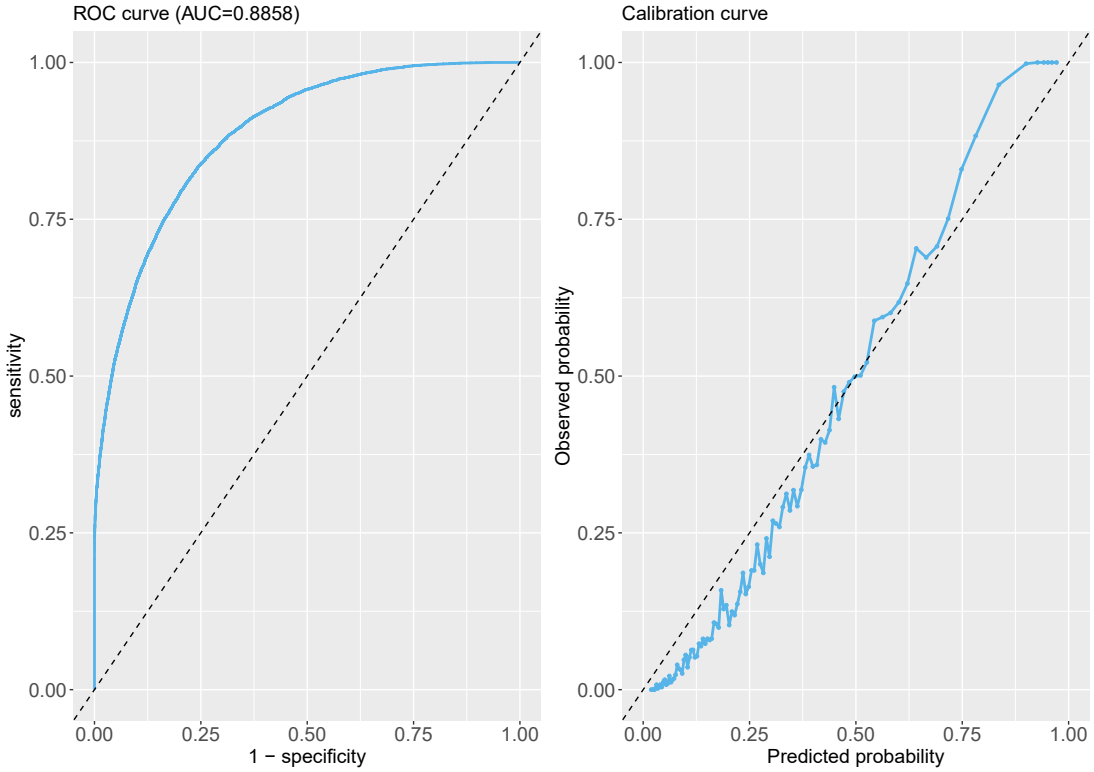



Figure 1.7: ROC and calibration curves of the GMERF.

outperform the CEP, EBP Log-Shift and FH logit in terms of RMSE. A direct comparison between GMERF and MERF shows that GMERF has better performance than MERF. These results are further supported by the discussion of mean and median values of bias and RRMSE in Table 1.5. GMERF displays the lowest bias for all 60 areas, while for the 52 in-sample areas, GMERF outperforms its competitors in terms of both mean and median bias.

Table 1.5: Mean and median of Bias and RRMSE over in- and out-of-sample areas for point estimates

| | Total | | In-sample | | Out-of-sample | |
|---|---|---|---|---|---|---|
| | Median | Mean | Median | Mean | Median | Mean |
| Bias | | | | | | |
| CEP | 0.0317 | 0.0301 | 0.0335 | 0.0373 | −0.0083 | −0.0170 |
| EBP Log-Shift | 0.0639 | 0.0611 | 0.0661 | 0.0695 | 0.0112 | 0.0068 |
| GMERF | 0.0282 | 0.0188 | 0.0288 | 0.0239 | −0.0065 | −0.0148 |
| MERF | 0.0343 | 0.0283 | 0.0370 | 0.0339 | −0.0158 | −0.0082 |
| FH logit | 0.0405 | 0.0389 | 0.0433 | 0.0444 | 0.0027 | 0.0030 |
| RRMSE[%] | | | | | | |
| CEP | 25.2968 | 25.9014 | 26.3609 | 27.1497 | 12.1399 | 17.7871 |
| EBP Log-Shift | 28.3263 | 28.9663 | 30.5159 | 30.4340 | 14.9920 | 19.4268 |
| GMERF | 16.6674 | 18.8713 | 16.6674 | 18.8498 | 10.4226 | 13.7361 |
| MERF | 21.6265 | 22.9136 | 22.5494 | 23.3961 | 11.6537 | 16.2025 |
| FH logit | 26.2628 | 27.7526 | 25.9965 | 26.9820 | 28.7644 | 32.7614 |

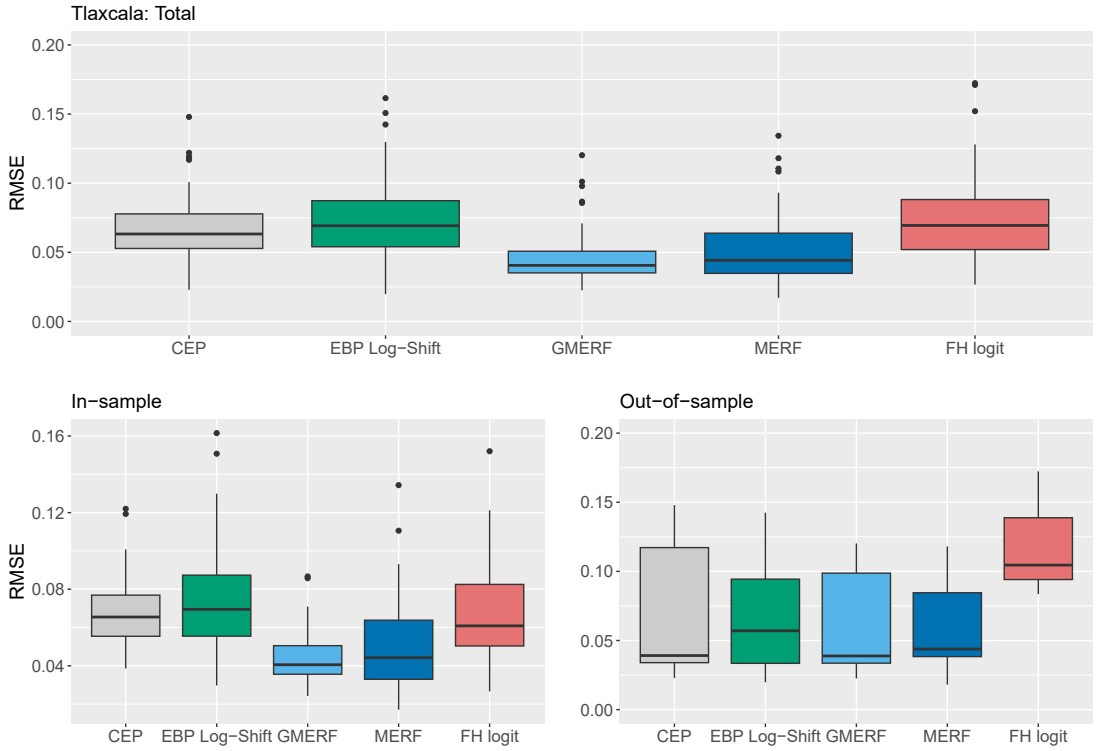



Figure 1.8: Performance of area-specific point estimates including details on in- and out-of-sample areas.

Lastly, we assess the performance of the proposed parametric MSE-bootstrap procedure. Table 1.6 presents the results for RB-RMSE and RRMSE-RMSE of the corresponding estimates. For in-sample areas, the median RB-RMSE suggests unbiasedness, while the mean RB-RMSE indicates a slight but acceptable level of overestimation. For out-of-sample areas, we observe a moderate level of overestimation leading to conservative MSE estimates. This is further supported by a moderate over-coverage of the nominal 95% confidence intervals, with a median CR of 97% across areas. Overall, the proposed parametric bootstrap procedure meets the expectations, given the challenging conditions of this design-based simulation.

Table 1.6: Performance of MSE-estimator in design-based simulation: mean and median of RB-RMSE and RRMSE-RMSE over in- and out-of-sample areas

| | Total | | In-sample | | Out-of-sample | |
|---|---|---|---|---|---|---|
| | Median | Mean | Median | Mean | Median | Mean |
| RB-RMSE[%] | −0.12 | 8.00 | −0.12 | 7.66 | 4.56 | 10.22 |
| RRMSE-RMSE[%] | 33.64 | 41.52 | 32.23 | 39.80 | 49.13 | 52.66 |

## 1.7 Conclusion

In this paper, we propose GMERFs for the estimation of disaggregated binary-based poverty indicators. Additionally, we investigate the impact of information loss resulting from the conversion of continuous variables into binary variables on the performance of estimation methods. Our results suggest that even when confronted with limited information on income, our method has the potential to deliver comparable performance to methods that necessitate more detailed income data on a continuous scale.

The proposed GMERF model employs a combination of the PQL method with an algorithm reminiscent of the EM algorithm, allowing for a flexible specification of the model. The resulting estimator for area-level proportions is complemented by a modified parametric bootstrap scheme similar to González-Manteiga et al. (2007). The performance of the point- and MSE estimates of the proposed method is evaluated against traditional SAE-methods for binary variables in a model-based simulation study. Furthermore, the proposed method is compared to established SAE-methods that use continuous household income as an input variable in an application and a design-based simulation study for estimating the HCR using income data from the Mexican state Tlaxcala. The model-based simulation demonstrates that our proposed point estimates perform well in scenarios with a linear specification and outperform traditional methods in the presence of non-linear interactions between covariates. The design-based simulation confirms the adequacy of GMERFs for point estimation under realistic conditions. We conclude that our proposed MSE-bootstrap scheme is reliable based on its performance in the model-based simulation, the application, and the design-based simulation.

From an applied perspective, additional research is required to establish a comprehensive metric indicating the extent to which SAE practitioners can forgo continuous income data and solely rely on rich or poor categorization to attain comparable estimators of point and uncertainty levels. A promising direction for future studies would involve extending the proposed approach to encompass count data. This expansion would enable the flexible estimation of a multidimensional poverty index, allowing for a more comprehensive assessment of poverty across multiple dimensions. Lastly, we propose a parametric bootstrap MSE for quantifying the uncertainty of the small area estimates. Exploring the development of non-parametric bootstrap and analytical estimators akin to those proposed by Krennmair and Schmid (2022) presents further prospects for future research. Finally, the proposed GMERF currently does not incorporate survey weights in the estimation, which poses a risk if the assumption of non-informative sampling is violated. This could potentially lead to biased estimates of area-level parameters. Therefore, extending the GMERF framework to account for informative sampling is a crucial direction for future research. For instance, the approaches proposed by Sverchkov and Pfeffermann (2018) and Parker et al. (2023) could be explored to adjust for the sampling design within the GMERF model. These methods typically involve either incorporating sampling weights directly into the estimation process or jointly modeling the sampling design and the target variable. Investigating such extensions would enhance the applicability of the GMERF method across a broader range of scenarios.

## Declarations

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

## 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. Finally, the authors are indebted to the Joint Editor, Associate Editor and two referees for comments that significantly improved the paper.

# Appendix A

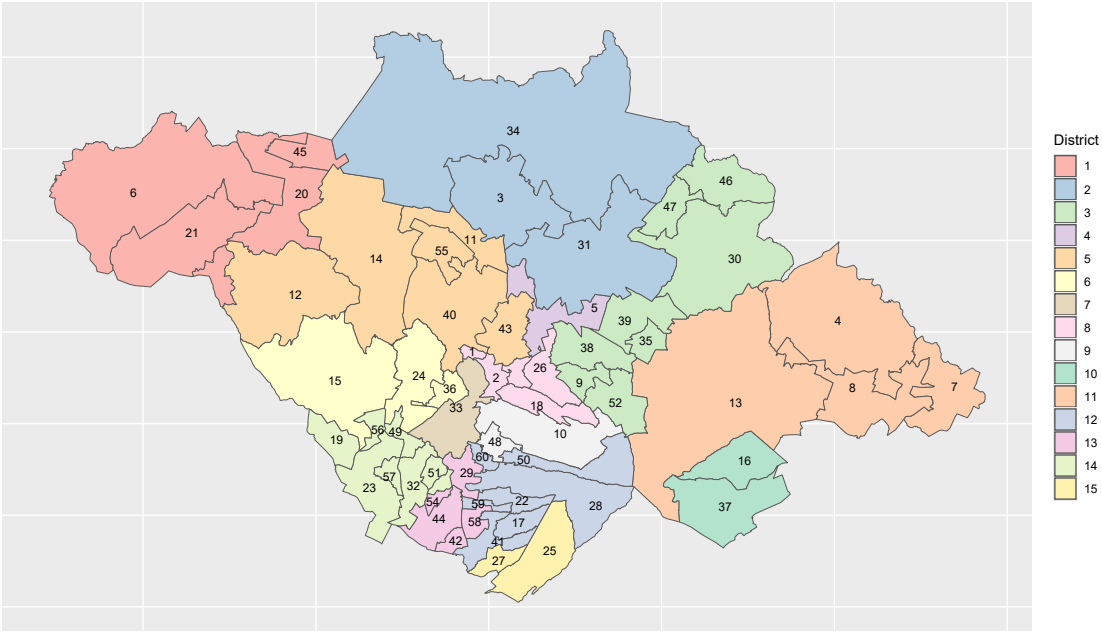



Figure A.1: Municipalities and districts in Tlaxcala.

Table A.1: List of the municipalities in Tlaxcala.

| ID | Municipality |
|---|---|
| 1 | Amaxac de Guerrero |
| 2 | Apetatitlán de Antonia Carvajal |
| 3 | Atlangatepec |
| 4 | Altzayanca |
| 5 | Apizaco |
| 6 | Calpulalpan |
| 7 | El Carmen Tequexquitla |
| 8 | Cuapiaxtla |
| 9 | Cuaxomulco |
| 10 | Chiautempan |
| 11 | Muñoz de Domingo Arenas |
| 12 | Españita |
| 13 | Huamantla |
| 14 | Hueyotlipan |
| 15 | Ixtacuixtla de Mariano Matamoros |
| 16 | Ixtenco |

*Continued on next page*

Table A.1 – *Continued from previous page*

| ID | Municipality |
|---|---|
| 17 | Mazatecochco de José Maria Morelos |
| 18 | Contla de Juan Cuamatzi |
| 19 | Tepetitla de Lardizábal |
| 20 | Sanctorum de Lázaro Cárdenas |
| 21 | Nanacamilpa de Mariano Arista |
| 22 | Acuamanala de Miguel Hidalgo |
| 23 | Nativitas |
| 24 | Panotla |
| 25 | San Pablo del Monte |
| 26 | Santa Cruz Tlaxcala |
| 27 | Tenancingo |
| 28 | Teolocholco |
| 29 | Tepeyanco |
| 30 | Terrenate |
| 31 | Tetla de la Solidaridad |
| 32 | Tetlatlahuca |
| 33 | Tlaxcala |
| 34 | Tlaxco |
| 35 | Tocatlan |
| 36 | Totolac |
| 37 | Zitlaltepec de Trinidad Sánchez Santos |
| 38 | Tzompantepec |
| 39 | Xaloztoc |
| 40 | Xaltocan |
| 41 | Papalotla de Xicohtencatl |
| 42 | Xicohtzinco |
| 43 | Yauhquemecan |
| 44 | Zacatelco |
| 45 | Benito Juárez |
| 46 | Emiliano Zapata |
| 47 | Lázaro Cárdenas |
| 48 | La Magdalena Tlaltelulco |
| 49 | San Damián Texoloc |
| 50 | San Francisco Tetlanohcan |
| 51 | San Jerónimo Zacualpan |
| 52 | San José Teacalco |
| 53 | San Juan Huactzinco |
| 54 | San Lorenzo Axocomanitla |
| 55 | San Lucas Tecopilco |
| 56 | Santa Ana Nopalucan |

*Continued on next page*

Table A.1 – *Continued from previous page*

| ID | Municipality |
|---|---|
| 57 | Santa Apolonia Teacalco |
| 58 | Santa Catarina Ayometla |
| 59 | Santa Cruz Quilehtla |
| 60 | Santa Isabel Xiloxoxtla |

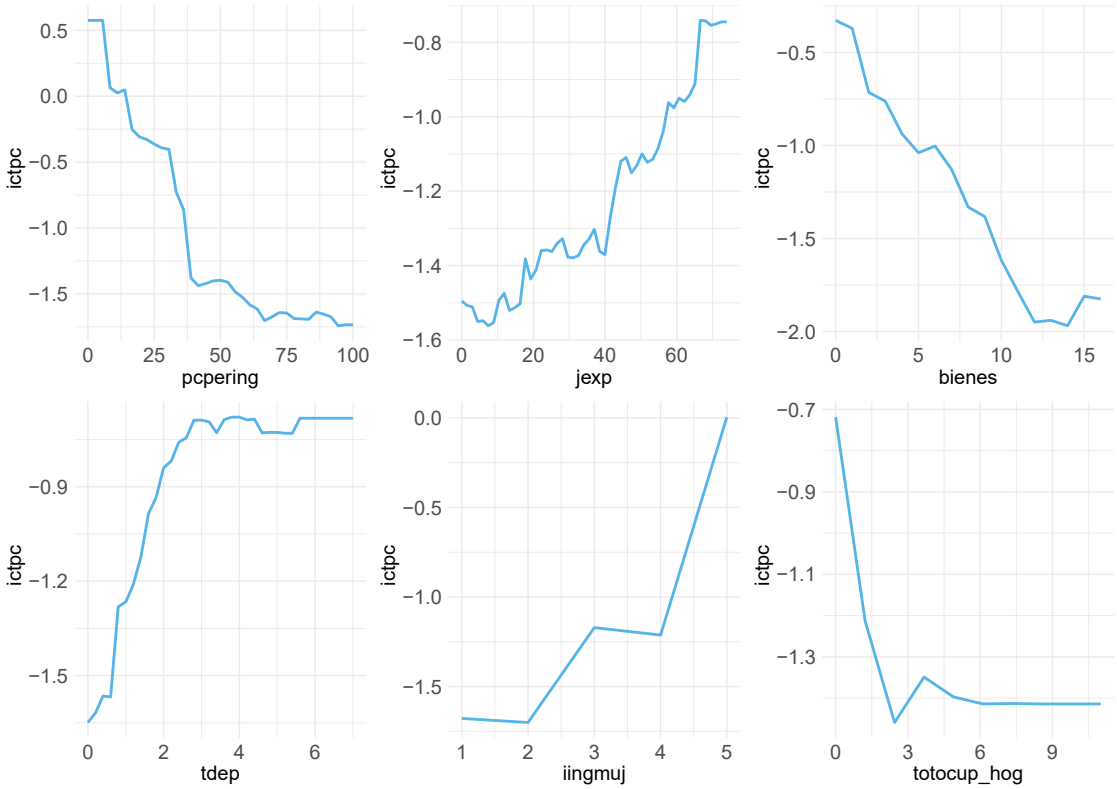



Figure A.2: Partial dependence plots: Predictive relations between dependent variable and covariates.

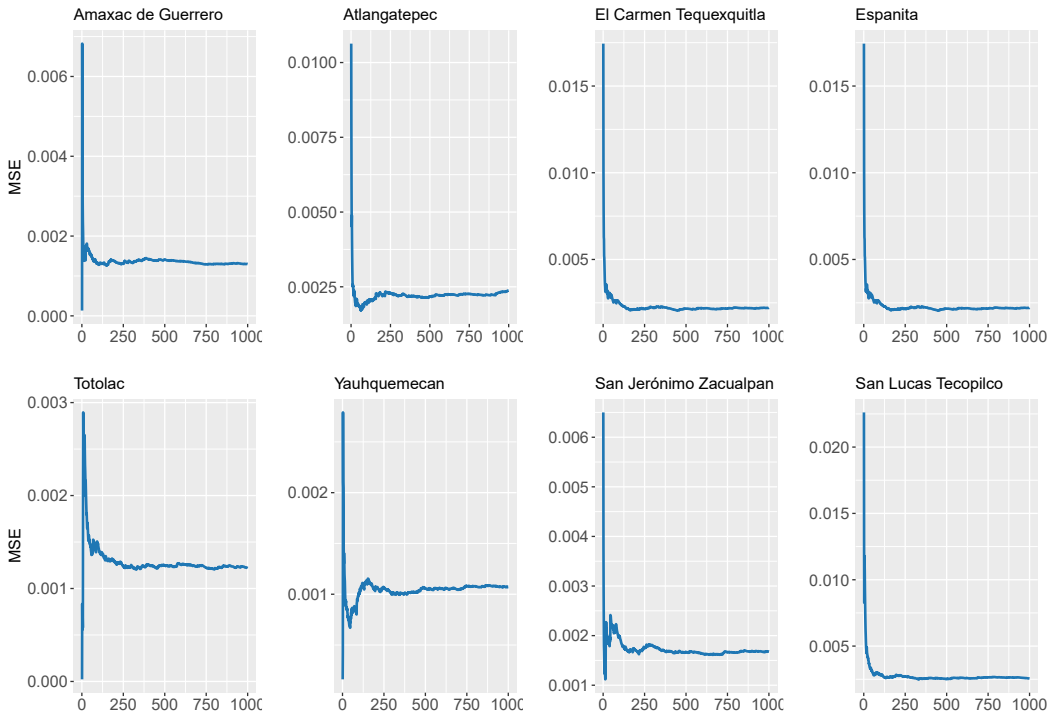



Figure A.3: Estimated MSE as a function of the number of bootstrap repetitions for the GMERF.

Table A.2: Description of the variables used in the EBP and FH models in the application presented in Section 1.5.

| Variable | Explanation | EBP | FH |
|---|---|---|---|
| *actcom* | Assets in the household | | ✓ |
| *autoconsumo* | Household with the presence of a member working in the primary sector | ✓ | |
| *bienes* | Availability of goods in the household | ✓ | |
| *ciclo_hog* | Household cycle | | ✓ |
| *clase_hog* | Household class | | ✓ |
| *id_men* | Households with minors under 16 years of age present | | ✓ |
| *iingmuj* | Identifies households with female-prevalence income | ✓ | |
| *jedad* | Age of the head of household | | ✓ |
| *jexp* | Years of working experience of the head of the household | ✓ | |
| *jnived* | Formal education of the head of the household | | ✓ |
| *jpea* | Occupational status of the head of household | ✓ | |
| *jubi* | Presence of retired people or pensioners in the household | ✓ | ✓ |
| *muj16_notrb_hog* | Presence in the household of women over 16 years of age not in employment | | ✓ |
| *pcpering* | Percentage of income earners in the household | ✓ | |
| *tam_hog* | Number of household members | | ✓ |
| *tdep* | Household members under 16 years and over 65, divided by the members between 16 to 64 | ✓ | |
| *totocup_hog* | Total number of employed in the household | ✓ | |
| *trabinf* | Households with paid child labor | | ✓ |

# Chapter 2

# Small area prediction of counts under machine learning-type mixed models

**Abstract**

Small area estimation methods are proposed that use generalized tree-based machine learning techniques to improve the estimation of disaggregated means in small areas using discrete survey data. Specifically, two existing approaches based on random forests - the Generalized Mixed Effects Random Forest (GMERF) and a Mixed Effects Random Forest (MERF) - are extended to accommodate count outcomes, addressing key challenges such as overdispersion. Additionally, three bootstrap methodologies designed to assess the reliability of point estimators for area-level means are evaluated. The numerical analysis shows that the MERF, which does not assume a Poisson distribution to model the mean behavior of count data, excels in scenarios of severe overdispersion. Conversely, the GMERF performs best under conditions where Poisson distribution assumptions are moderately met. In a case study using real-world data from the state of Guerrero, Mexico, the proposed methods effectively estimate area-level means while capturing the uncertainty inherent in overdispersed count data. These findings highlight their practical applicability for small area estimation.

**Keywords**: Bootstrap, Generalized linear mixed models, Overdispersed count data, Random forest, Small area estimation

## 2.1 Introduction

Linear mixed models (LMM) are a staple for analyzing unit-level survey data and estimating area-level means. These models account for the hierarchical structure of observations through random effects. An example of such a model is the nested error regression model (Battese et al., 1988), which requires the availability of unit-level survey data and administrative auxiliary information, such as data from a register or census. However, it is crucial to recognize the limitations imposed by the distributional and structural assumptions of these models, which may not always hold in small area estimation (SAE) applications (Rao and Molina, 2015). For example, when working with LMMs, it is necessary to assume a linear relationship between the covariates and the outcome variable. This assumption may not always align with empirical evidence. Thus, it is critical to ensure the validity of model assumptions for optimal results and predictive performance of model-based SAE. If assumptions are not met, parameter estimates may be biased, and the reliability of mean squared error (MSE) estimates may be compromised

(Jiang and Rao, 2020).

To circumvent the parametric assumptions inherent in LMMs, employing machine learning techniques is a viable methodological alternative. These techniques are capable of extracting predictive relationships from data, encompassing complex interactions among covariates, without relying on explicit model assumptions (Hastie et al., 2009; Varian, 2014). Krennmair and Schmid (2022) provide a framework utilizing tree-based machine learning methods for SAE. This framework presents a non-linear, data-driven, and semi-parametric approach for continuous variables to estimate area-level means using Mixed Effects Random Forests (MERF). Random forests (Breiman, 2001) are renowned for their robust predictive performance, robustness with respect to model-misspecification, and they inherently address model-selection issues (Biau and Scornet, 2016). MERFs incorporate these benefits while also modeling hierarchical dependencies. Originally developed for continuous target variables similar to LMMs, Frink and Schmid (2024a) have expanded the application of MERFs to estimate poverty indicators using binary response variables within the context of SAE. However, their focus is limited to binary variables, and the behavior of MERFs with count data remains unexplored.

Various methods for SAE with count data have been investigated. Parametric models are discussed by Ghosh et al. (1998), Jiang and Lahiri (2006), Dreassi et al. (2014), Chen et al. (2015), Boubeta et al. (2016), Hobza and Morales (2016), Boubeta et al. (2017) and Chandra et al. (2017). In contrast, Chambers et al. (2014) and Tzavidis et al. (2015) have examined semi-parametric unit-level models utilizing M-quantiles, which only partially rely on the distributional assumptions of the Poisson and Negative Binomial families. However, all of these methodologies for count data assume a linear relationship within the systematic component of the underlying model.

Building on the work of Krennmair and Schmid (2022) and Frink and Schmid (2024a), who originally developed the (generalized) MERF approaches for continuous and binary data, this paper introduces semi-parametric small area estimation methods for count data using random forests. These extensions allow for a more flexible specification of the relationship between the outcome variable and the covariates, enabling their application to scenarios involving count outcomes.

The generalized MERF (GMERF) is presented as a flexible and data-driven approach that employs the Poisson distribution to model the mean behavior of count data. Furthermore, a parametric MSE bootstrap scheme is proposed to evaluate the uncertainty associated with area-level estimates. While commonly used for modeling count data, the Poisson distribution often underestimates variability in cases of overdispersion (Ver Hoef and Boveng, 2007). To address this challenge, we introduce a semi-parametric MSE bootstrap procedure. This bootstrap scheme for the GMERF operates with reduced reliance on distributional assumptions, providing a safeguard against mild deviations from the Poisson distribution. The numerical findings from our simulation studies and real-world application indicate that GMERF performs best when the assumptions of the Poisson distribution are moderately met, particularly in scenarios with minimal to low overdispersion and non-linear interactions between covariates.

Although the original MERF approach is designed for continuous outcome variables, we extend it to accommodate count data. Our numerical analysis demonstrates that the MERF pro-

vides a valid alternative for modeling count data, effectively capturing complex relationships and handling severe overdispersion without relying on explicit distributional assumptions like Poisson or Negative Binomial. Furthermore, we evaluate its numerical performance in comparison to GMERF through simulation studies. Additionally, we introduce a modified non-parametric bootstrap method for MERF, inspired by Krennmair and Schmid (2022), to account for the additional uncertainty that arises when treating a count variable as continuous.

The paper is organized as follows: Section 2.2.1 introduces GMERFs as a methodology that merges random forests with hierarchical modeling to capture dependencies among unit-level (count) observations. Section 2.2.2 details the process for constructing area-level estimates. To accurately evaluate the MSE of these estimates, Section 2.3 describes three bootstrap schemes: parametric and semi-parametric approaches for GMERFs, and an adjusted non-parametric method for the MERF. Section 2.4 assesses and compares the effectiveness of the proposed GMERF and MERF methods against the empirical best plug-in predictor (EBPP, Jiang (2003)) under the Poisson generalized linear mixed model through model-based simulation studies. Section 2.5 applies these methods to estimate area-level means of educational attainment among women in the Mexican state of Guerrero, also addressing the uncertainty of these estimates. Section 2.6 presents a design-based simulation study, based on data from Guerrero, to assess the proposed estimators in a close-to-reality environment. The study concludes in Section 2.7 with recommendations for further research in the field of SAE.

## 2.2 Methodology

Krennmair and Schmid (2022) introduced the MERF for continuous data in the context of SAE. This section introduces an extension of the MERF, the GMERF, a semi-parametric unit-level model designed to handle Quasi-Poisson-distributed count outcomes. The model employs a regression forest approach to estimate area-level means, utilizing both unit-level survey and unit-level administrative data. This technique effectively addresses the challenges of modeling count data by combining the flexibility of machine learning methods, which can account for non-linear relationships, with parametric Poisson assumptions.

### 2.2.1 Generalized semi-parametric unit-level model

We examine a finite population $P$ divided into $D$ disjunct areas $P_i$ each with a sub-population size of $N_i$, 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 a total size $n = \sum_{i=1}^{D} n_i$. In contrast, non-sampled observations are denoted as $r_i$ with a size of $N_i - n_i$. We denote individual units within each area as $j \in s_i$ for sampled and $j \in r_i$ for non-sampled observations. The discrete unit-level outcome variable is represented by $y_{ij}$, with information available for $n$ observations within the sample $s$. The vector $\mathbf{x}_{ij} = [x_1, x_2, ..., x_p]^\intercal$ encompasses $p$ auxiliary covariates, and these auxiliary variables are known for all $N$ units in our population $P$. We assume that $y_{ij}$ follows a generalized semi-parametric unit-level model:

$$\eta_{ij} = f(\mathbf{x}_{ij}) + \nu_i, \tag{2.1}$$

$$\nu_i \sim N(0, \sigma_\nu^2),$$
$$E(y_{ij}|\nu_i) = \mu_{ij} = g(\eta_{ij}),$$

where the fixed part, $f()$, is defined by a random forest and models the conditional mean of the linear predictor ($\eta_{ij}$) given the covariates. The term $\nu_i$ denotes a domain-specific random intercept with its variance $\sigma_\nu^2$, capturing small-area variations in the conditional distribution of the linear predictor given $\mathbf{x}_{ij}$. $g()$ is a monotonic, differentiable link function that specifies the function of the mean ($\mu_{ij}$) equated to the systematic component. In this work, we assume that the individual $y_{ij}$ values in domain $i$ are independent Poisson random variables with $Var(y_{ij}|\nu_i) = E(y_{ij}|\nu_i)$. Consequently, the mean of the response variable is connected to the linear predictor via the logarithmic link function: $\mu_{ij} = \exp(\eta_{ij})$. If $y_{ij}$ were a continuous variable following a Gaussian distribution, the link function would simplify to the identity function. Thus, in model (2.1), the GMERF reduces to the MERF, as proposed by Krennmair and Schmid (2022). Furthermore, model (2.1) can be seen as a generalized linear mixed model (GLMM) by setting $f(\mathbf{x}_{ij}) = \mathbf{x}_{ij}^\intercal \beta$, where $\beta = [\beta_1, \ldots, \beta_p]^\intercal$ represents the regression parameters.

Given that inference using GLMMs presents computational hurdles due to the presence of high-dimensional integrals in the likelihood, which are not amenable to analytical evaluation, we employ an estimation strategy for model (2.1) akin to Frink and Schmid (2024a). The key distinction in our approach lies in the use of a different link function, tailored to the count data in our model, while the other elements of the estimation strategy align closely with those described by Frink and Schmid (2024a). Specifically, this approach utilizes the penalized quasi-likelihood (PQL) method (Breslow and Clayton, 1993; Stroup, 2012) in conjunction with the EM-algorithm (Moon, 1996) to address the computational challenges inherent in GLMM estimation. More concretely, a weighted MERF pseudo-model is devised, leveraging a linearized target variable $y_{L,ij} = g(\mu_{ij}) + (y_{ij} - \mu_{ij})g'(\mu_{ij})$ and weights $w_{ij} = (v_{ij}g'(\mu_{ij})^2)^{-1}$, with $v_{ij}$ being a known variance function. The algorithm follows a doubly iterative process, with micro iterations nested within macro iterations. During each macro-iteration, the linearized response variable and weights are updated. These revised values then serve as the response variable and weights for the subsequent micro-iterations. To fit model (2.1) on survey data, the GMERF algorithm uses initial estimates for $\mu_{ij}$, weights $w_{ij}$ and $y_{L,ij}$ from a generalized linear model (GLM) during the first macro-iteration. In the subsequent micro-iterations, the algorithm (i) estimates the forest function, assuming the random effects to be correct, and (ii) estimates the random effects part using the weighted pseudo-model, assuming the Out-of-Bag (OOB) predictions from the regression forest to be valid. Convergence is monitored by the relative change in the log-likelihood of model (2.1) between two micro-iterations. It is considered achieved when its relative change is less than a specified number. Once convergence is achieved, the initial values for the next macro-iteration are updated. For the specific case of $y_{ij}$ being Poisson distributed, the initial values are updated as follows:

$$\eta_{ij} = \hat{f}(\mathbf{x}_{ij})^{OOB} + \hat{\nu}_i, \tag{2.2}$$
$$\mu_{ij} = \exp(\eta_{ij}),$$

$$y_{L,ij} = \log(\mu_{ij}) + \frac{(y_{ij} - \mu_{ij})}{\mu_{ij}},$$
$$w_{ij} = \frac{1}{v_{ij}(\mu_{ij})},$$

where $\hat{f}(\mathbf{x}_{ij})^{OOB}$ and $\hat{\nu}_i$ reflect their estimated values from the micro-level convergence achieved in the preceding macro-iteration. The predictive performance of our method is significantly influenced by two key tuning parameters: the number of split candidates at each node, which controls the degree of decorrelation, and the number of trees, which together enhance the model's stability and accuracy. For further methodological details, we refer to Frink and Schmid (2024a).

The Poisson distribution is very common for modeling the mean behavior of count outcomes. However, it may underestimate variability in cases of overdispersion, where the observed variance of the response variable exceeds what the Poisson distribution predicts. Overdispersion is a common occurrence in count data analysis and can significantly affect result interpretation. Failure to account for overdispersion in the model can result in overly narrow standard errors and confidence intervals, as well as excessively lenient significance tests, leading to the detection of effects that do not truly exist (Ver Hoef and Boveng, 2007). One widely used approach to address overdispersion is to employ a Quasi-Poisson model (Gourieroux et al., 1984). In Quasi-Poisson models, $E(y_{ij}^{qp}|\nu_i) = E(y_{ij}|\nu_i) = \mu_{ij}$ and $Var(y_{ij}^{qp}|\nu_i) = \theta\mu_{ij}$, where $\theta$ is a dispersion parameter. The algorithm mentioned above can be directly expanded to incorporate Quasi-Poisson models to address overdispersion concerns. Nevertheless, given our primary focus on predicting means within the machine learning framework and our lesser concern with standard error inference, the Quasi-Poisson model does not yield any discernible advantage within the generalized mixed effect regression forest model. Results for the Quasi-Poisson distribution in terms of GMERFs are available and can be requested from the authors.

### 2.2.2 Domain-level estimator for counts

The proposed estimator for the area-level mean is given by:

$$\hat{\eta}_{ij} = \hat{f}(\mathbf{x}_{ij}) + \hat{\nu}_i,$$
$$\hat{\mu}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \exp(\hat{\eta}_{ij}) \text{ for } i = 1, \ldots, D. \tag{2.3}$$

For non-sampled areas, the proposed estimator for the area-level mean simplifies to the fixed component obtained from the random forest:

$$\hat{\eta}_{ij} = \hat{f}(\mathbf{x}_{ij}),$$
$$\hat{\mu}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \exp(\hat{\eta}_{ij}).$$

Our proposed estimator $(2.3)$ offers an advancement over existing SAE methods for count data by addressing issues commonly associated with model selection. This improvement arises from

employing the random forest technique, which inherently optimizes model selection to capture higher-order effects and non-linear interactions. This feature is practically demonstrated in our application to real-world data from the state of Guerrero in Section 2.5. Furthermore, the random forest is well-suited for analyzing high-dimensional covariate data, effectively managing scenarios where the number of covariates exceeds the sample size (Hastie et al., 2009).

However, the GMERF estimator relies on the assumptions of the Poisson distribution due to the linearized target variable and the weights used in the algorithm. In contrast, the MERF, originally designed for continuous outcomes, offers a viable alternative for modeling count data. This is achieved by having the MERF iteratively estimate a random forest, assuming the random effects term to be correct, and then estimating the random effects components (and variance components) by decomposing the random forest residuals using a linear mixed model. Moreover, the MERF has the advantage of not depending on the assumptions of the Poisson or Negative Binomial distributions, making it an alternative option when dealing with severe overdispersion.

## 2.3 Uncertainty estimation

Estimating the MSE of small area estimates poses a substantial challenge (Rao and Molina, 2015). In this section, we present an adaption of two existing bootstrap schemes, proposed by González-Manteiga et al. (2007) and Chambers and Chandra (2013), to estimate the MSE of the small area estimator introduced in equation (2.3). The introduced bootstrap methods account for the additional uncertainty associated with modeling count outcomes. The primary distinction between the two bootstrap procedures lies in the mechanism used to generate the bootstrap population. Following González-Manteiga et al. (2007) and González-Manteiga et al. (2008), the first bootstrap scheme generates bootstrap realizations of the random effects parametrically and explicitly employs the Poisson distribution to simulate $y_{ij}$. This approach is widely used in SAE applications, such as in Boubeta et al. (2016) for Poisson mixed models. It is particularly effective in scenarios where the target variable follows a Poisson distribution, as it directly leverages the properties of this distribution. However, count data often exhibit overdispersion, which violates the assumptions underlying the Poisson distribution. As a result, the parametric bootstrap may become unsuitable in such cases. To address this, we propose adopting a semi-parametric bootstrap approach for the GMERF - at least as a supplementary MSE estimation method - drawing on the work of Chambers and Chandra (2013), Flores-Agreda and Cantoni (2019) and Krennmair and Schmid (2022). Since this approach is less reliant on the assumptions of the Poisson distribution, we expect the semi-parametric bootstrap to serve as a safeguard against deviations from the Poisson distribution, particularly in the presence of overdispersion.

Furthermore, we introduce an adjusted non-parametric MSE bootstrap scheme for the MERF with count outcomes. Building on the work of Chambers and Chandra (2013) and Krennmair (2022), this bootstrap procedure is adjusted to account for the uncertainty of the estimation, enabling the generation of a discrete count variable in the bootstrap population. Detailed information on this bootstrap method can be found in Appendix B.1.

### 2.3.1 Parametric bootstrap

The procedural steps outlined for the proposed parametric bootstrap of the GMERF closely resemble those detailed in Frink and Schmid (2024a) and can be summarized as follows:

1. For $b = 1, \ldots, B$:
   (a) Create bootstrap random effects for each of the $D$ areas by $\nu_i^{(b)} \sim N(0, \hat{\sigma}_\nu^2)$.
   (b) Create a bootstrap population of size $N$, by generating values $y_{ij}^{(b)}$ from a Poisson distribution with
   $$\mu_{ij}^{(b)} = \exp\left(\hat{f}(\mathbf{x}_{ij}) + \nu_i^{(b)}\right).$$
   (c) Determine the true bootstrap population area means $\mu_i^{(b)}$ as $\frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}^{(b)}$ for all $i = 1, \ldots, D$.
   (d) For each bootstrap population, select a bootstrap sample that matches the original sample size $n_i$ and estimate $\hat{f}^{(b)}()$ and $\hat{\nu}_i^{(b)}$. Calculate area-level means $\hat{\mu}_i^{(b)}$.
2. Utilizing the $B$ bootstrap samples, the MSE estimator is derived as follows:
   $$\widehat{MSE}_i = \frac{1}{B}\sum_{b=1}^{B}\left(\mu_i^{(b)} - \hat{\mu}_i^{(b)}\right)^2.$$

### 2.3.2 Semi-parametric bootstrap

The parametric bootstrap may be unsuitable in the presence of overdispersion. To address this, we discuss a semi-parametric bootstrap approach that reduces reliance on strict Poisson distribution assumptions. This bootstrap scheme for the GMERF ensures that a discrete count target variable is generated in the bootstrap population without explicitly relying on the Poisson distribution. The proposed bootstrap scheme is described below:

1. For given $\hat{f}()$, calculate the marginal pearson residuals $z_{ij} = \frac{y_{ij} - \exp(\hat{f}(\mathbf{x}_{ij}))}{\sqrt{\exp(\hat{f}(\mathbf{x}_{ij}))}}$.
2. Employing the marginal pearson residuals $z_{ij}$, calculate 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. Compute the vector of level-1 residuals by $\hat{z}_{ij} = z_{ij} - \bar{z}_i$. Similar to Krennmair and Schmid (2022), we adjust the residuals $\hat{z}_{ij}$ to the bias-corrected variance and center them, indicated by $\hat{z}_{ij}^c$. Our approach differs in that we use the marginal Pearson residuals of the GMERF, as described in step 1 above, instead of the marginal raw residuals of the MERF for the adjustment process. Level-2 residuals $\bar{z}_i$ are additionally adjusted for the estimated variance $\hat{\sigma}_\nu^2$ and centered, indicated by $\bar{z}^c = (\bar{z}_1^c, ..., \bar{z}_D^c)$.
4. For $b = 1, ..., B$:

(a) Independently draw samples with replacement from the scaled and centered level-1 and level-2 residuals:

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

(b) Calculate $\mu_{ij}^{(b)} = \exp(\hat{f}(\mathbf{x}_{ij}) + \bar{z}_{i}^{(b)})$ and the corresponding $\tilde{y}_{ij}^{(b)} = \mu_{ij}^{(b)} + \sqrt{\mu_{ij}^{(b)}} \times z_{ij}^{(b)}$ for $j = 1, \ldots, N$. Note that $\tilde{y}_{ij}^{(b)}$ is defined on a continuous scale. Following Flores-Agreda and Cantoni (2019), the inclusion of the residuals $z_{ij}^{(b)}$ ensures that the bootstrap data capture the variability observed in the original data, thereby preventing the underestimation of uncertainty. Scaling these residuals by the estimated standard deviation $\sqrt{\mu_{ij}^{(b)}}$ helps to maintain consistency in variability between the bootstrap and original data.

(c) To obtain a count target variable, we match $\tilde{y}_{ij}^{(b)}$ with the set of estimated unit-level predictors from the sample $\left\{\hat{\mu}_t |\ \hat{\mu}_t = \exp(\hat{f}(\mathbf{x}_{it}) + \hat{\nu}_i),\ \text{for}\ t = 1, ..., n\right\}$ by finding the corresponding index $\tilde{t}$ solving

$$\min_{t} \left|\tilde{y}_{ij}^{(b)} - \hat{\mu}_t\right|.$$

(d) Define the bootstrap population by $y_{ij}^{(b)} = y_{\tilde{t}}$ for $j = 1, \ldots, N$ and calculate the true bootstrap population mean $\mu_{i}^{(b)}$ for $i = 1, ..., D$.

(e) For each bootstrap population, select a bootstrap sample that matches the original sample size $n_i$ and estimate $\hat{f}^{(b)}()$ and $\hat{\nu}_{i}^{(b)}$. Obtain estimates for the mean $\hat{\mu}_{i}^{(b)}$.

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

$$\widehat{MSE}_i = \frac{1}{B}\sum_{b=1}^{B}\left(\mu_{i}^{(b)} - \hat{\mu}_{i}^{(b)}\right)^2.$$

Both bootstrap schemes we described in this section are empirically evaluated in Section 2.4.

## 2.4 Model-based simulation study

The use of model-based simulations enables a controlled empirical assessment of our proposed methods for point and uncertainty estimates. We compare the point and uncertainty estimates for domain-level means derived from the GMERF model (2.1) with those obtained from two competing models. In particular, we examine the performance of GMERFs in comparison to the EBPP, which is based on a Poisson GLMM:

$$\hat{\mu}_{i}^{EBPP} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{\mu}_{ij}\right),$$

where $\hat{\mu}_{ij} = \exp(\mathbf{x}_{ij}^{\intercal}\hat{\beta} + \hat{\nu}_i)$ (Jiang, 2003). By contrasting the performance of this linear-based competitor (on the linear predictor sale) with our more flexible approach, which incorporates

semi-parametric and non-linear modeling, we aim to showcase the advantages of the GMERF methodology. Additionally, we compare the GMERF method with the MERF (for originally continuous outcomes):

$$\hat{\mu}_i^{MERF} = \frac{1}{N_i} \sum_{j=1}^{N_i} \hat{f}(\mathbf{x}_{ij}) + \hat{\nu}_i.$$

The aim of comparing both machine learning methods is to highlight the advantages of GMERF when the Poisson distribution is applicable or only moderately violated, and to demonstrate the suitability of MERF for count data in situations of severe overdispersion.

The simulation setup involves a finite population $U$ with a total size of $N = 50,000$, divided into $D = 50$ disjoint areas $U_1, ..., U_D$ each having $N_i = 1,000$ units. We generate samples using stratified random sampling, where the $50$ small areas are treated as strata. This approach yields a total sample size of $n = \sum_{i=1}^{D} n_i = 921$. The number of sampled units per area varies from $8$ to $29$, with a median of $18$. These sample sizes are consistent with the area-level sample sizes found in the application discussed in Section 2.5.

We consider four scenarios denoted as *Normal-Poisson*, *Interaction-Poisson*, *NB3*, *NB1* and repeat each scenario independently $M = 500$ times. The comparison of competing model-estimates under these four scenarios allows us to examine the performance under two major dimensions: Firstly, the presence of overdispersed data delineated by the Negative Binomial distribution and secondly, the presence of unknown non-linear interactions between covariates on the linear predictor scale.

We initially establish a baseline scenario, *Normal-Poisson*, for Poisson GLMMs generating Poisson-distributed outcomes without interactions or quadratic terms. Since the assumption of linearity in the model is satisfied, our goal is to demonstrate that GMERFs perform similar to their linear counterparts in the reference scenario. In contrast, the *Interaction-Poisson* scenario introduces a more intricate model, incorporating quadratic terms and interactions at the linear predictor level. This scenario aims to underscore the benefits of non-linear modeling approaches. Overdispersion frequently complicates the analysis of count data. To more realistically mirror these situations, we apply a Negative Binomial distribution (with scale parameter $3$ and $1$) in scenarios *NB3* and *NB1*. These scenarios utilize complex models with interaction effects but vary according to the scale parameter $s$, which influences the degree of overdispersion. A smaller $s$ value indicates increased overdispersion. Further information on the data generation process for each scenario is detailed in Table 2.1.

We evaluate the point estimates for the area-level means using two quality measures: bias and root mean squared error (RMSE). To assess the proposed MSE estimators, we examine the relative bias of the root mean squared error (RB-RMSE) and the relative root mean squared error of the RMSE:

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

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

Table 2.1: Model-based simulation scenarios.

| Scenario | Linear predictor | Mean | $y$ | $x_1$ | $x_2$ | $\nu$ |
|---|---|---|---|---|---|---|
| *Normal-Poisson* | $\eta = 2 + x_1 + x_2 + \nu$ | $\mu = \exp(\eta)$ | Pois$(\mu)$ | $U(-1,1)$ | $N(-1,1)$ | $N(0, 0.3^2)$ |
| *Interaction-Poisson* | $\eta = 2 + 2x_1x_2 + x_2^2 + \nu$ | $\mu = \exp(\eta)$ | Pois$(\mu)$ | $U(-1,1)$ | $N(-1,1)$ | $N(0, 0.3^2)$ |
| *NB3* | $\eta = 2 + 2x_1x_2 + x_2^2 + \nu$ | $\mu = \exp(\eta)$ | NB$(\mu, 3)$ | $U(-1,1)$ | $N(-1,1)$ | $N(0, 0.3^2)$ |
| *NB1* | $\eta = 2 + 2x_1x_2 + x_2^2 + \nu$ | $\mu = \exp(\eta)$ | NB$(\mu, 1)$ | $U(-1,1)$ | $N(-1,1)$ | $N(0, 0.3^2)$ |

$$RB\text{-}RMSE_i = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M} MSE_{est_i}^{(m)}} - RMSE_i}{RMSE_i}$$

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

where $\hat{\mu}_i^{(m)}$ denotes the estimated mean for area $i$ derived from any of the aforementioned methods, and $\mu_i^{(m)}$ represents the true mean for area $i$ in simulation round $m$. The estimation of $MSE_{est_i}^{(m)}$ is carried out using the proposed bootstrap methods detailed in Section 2.3 and Appendix B.1.

To conduct the model-based simulation, we utilize R (R Core Team, 2024). The EBPP estimates are generated using the **lme4** package (Bates et al., 2015). For the proposed GMERF and MERF methods, we use the **ranger** package (Wright and Ziegler, 2017) alongside **lme4** (Bates et al., 2015). To ensure the algorithms converge, we apply a precision of $1e^{-5}$ for the relative change in the log-likelihood criterion (applicable to the MERF algorithm as well) and a precision of 0.001 for the relative change in $\hat{\eta}$.

Table 2.2 reports the empirical bias and the RMSE of each method across the four scenarios. In the *Normal-Poisson* scenario the EBPP, MERF, and GMERF estimators show varying degrees of positive bias, with the median bias magnitude increasing in that order. For the *Interaction - Poisson* data-generating process, biases not only increase but also change sign for all methods, indicating more complex error dynamics. In the Negative Binomial scenarios, designed to examine model performance under conditions of overdispersion, the MERF demonstrates a lower bias in comparison to the GMERF. The GMERF shows the highest bias across these scenarios, likely due to its reliance on the approximation method in its algorithm.

Table 2.2: Mean and median of bias and RMSE over areas for point estimates.

| | *Normal-Poisson* | | *Interaction-Poisson* | | *NB3* | | *NB1* | |
|---|---|---|---|---|---|---|---|---|
| | Median | Mean | Median | Mean | Median | Mean | Median | Mean |
| Bias | | | | | | | | |
| EBPP | 0.0054 | 0.0123 | −0.0895 | −0.0887 | −0.0609 | −0.0489 | −0.0894 | −0.1013 |
| GMERF | 0.0774 | 0.0834 | −0.1376 | −0.1587 | 0.4432 | 0.4981 | 3.1600 | 3.1510 |
| MERF | 0.0075 | 0.0099 | −0.0628 | −0.0628 | 0.0830 | 0.0592 | 0.0675 | 0.0750 |
| RMSE | | | | | | | | |
| EBPP | 1.0910 | 1.1525 | 1.5450 | 1.6360 | 4.3740 | 4.5650 | 7.3670 | 7.5960 |
| GMERF | 1.2449 | 1.2979 | 1.4300 | 1.4860 | 4.2140 | 4.3490 | 8.2460 | 8.8140 |
| MERF | 1.3610 | 1.4370 | 1.9300 | 2.0120 | 4.1360 | 4.2170 | 5.6160 | 5.6930 |

Figure 2.1, along with Table 2.2, evaluates the RMSE for each estimation method across various scenarios. In the *Normal-Poisson* scenario, EBPP outperforms both MERF and GMERF. This suggests that EBPP's adherence to the model's fixed effects leads to more accurate estimates when the assumptions of the Poisson distribution are met. In the more complex scenario *Interaction-Poisson*, MERF shows a higher RMSE compared to EBPP, indicating lower accuracy. MERF's performance suggests a greater sensitivity to deviations of distribution assumptions compared to violations of linearity assumptions, as EBPP remains more efficient under the Poisson assumption with non-linear relationships in the fixed effects part of the model. Notably, GMERF's point estimates surpass those of EBPP, highlighting its effectiveness in complex models that involve interactions and non-linear relationships. In scenarios incorporating a Negative Binomial distribution, MERF's performance tends to improve as the degree of overdispersion increases. This superiority (especially in the *NB1* scenario) highlights MERF's ability to effectively handle overdispersed count data, benefiting from its flexibility in modeling both continuous and discrete targets without strict reliance on Poisson model assumptions. Overall, the insights from Table 2.2 and Figure 2.1 indicate that GMERF delivers competitive performance when the Poisson distribution assumptions are moderately met, whereas MERF excels in scenarios where these assumptions are severely violated.

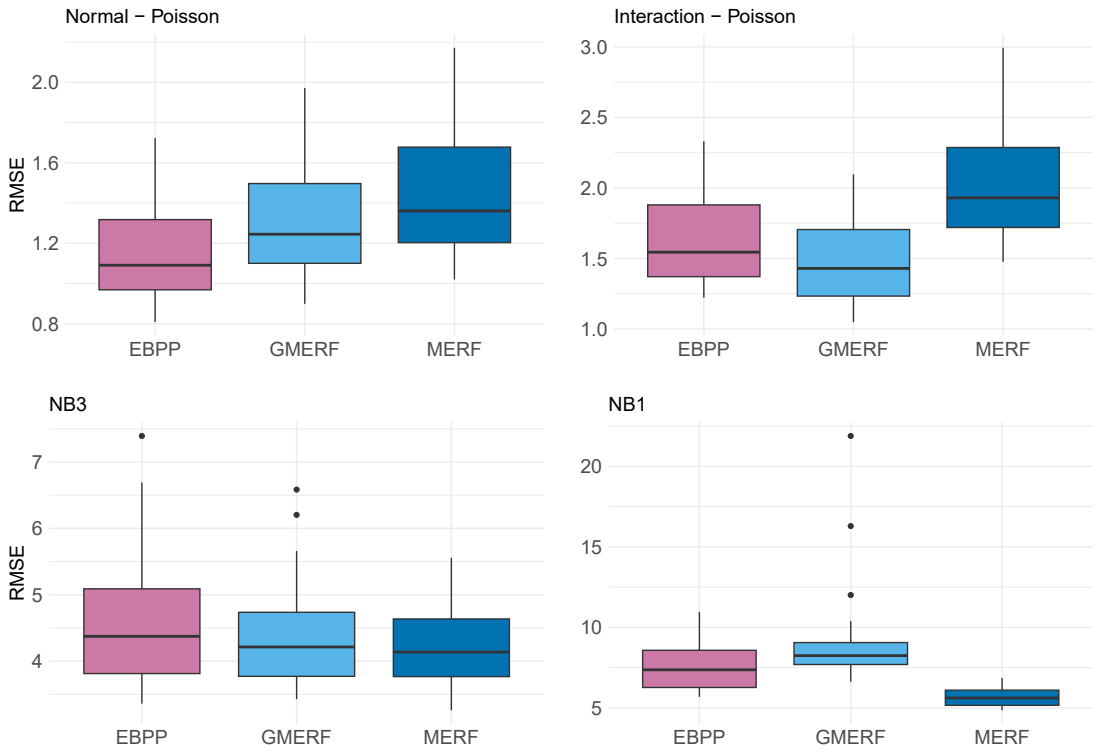



Figure 2.1: Empirical RMSE comparison of point estimates for area-level averages under four scenarios.

We now assess the performance of the MSE estimators presented in Section 2.3 and Appendix B.1 using $B = 200$ bootstrap replications. Table 2.3 displays the performance of the three bootstrap procedures, evaluated using RB-RMSE and RMSE-RMSE. Specifically, the proposed MSE estimators show reasonably low to moderate relative bias in terms of both

mean and median values across all scenarios. The parametric bootstrap estimator (GMERF P) performs similarly in simpler scenarios, such as *Normal-Poisson*, and in more complex scenarios like *Interaction-Poisson*. The results for Negative Binomial settings using this parametric bootstrap are affected by its reliance on Poisson distribution assumptions, which can lead to severe underestimation when these assumptions are violated. The semi-parametric bootstrap estimator for GMERF (GMERF SP) exhibits slight variability in relative bias across different scenarios. The adjusted non-parametric bootstrap estimator for MERF (MERF NPC) consistently shows low bias and maintains precision across various scenarios. Notably, GMERF SP performs comparably to MERF NPC in scenarios involving a Negative Binomial distribution in terms of RB-RMSE. While Table 2.3 does not directly convey the area tracking properties of the estimated RMSE versus the empirical RMSE, Figure 2.2 provides further insights. Based on the tracking properties, we conclude that using the parametric, semi-parametric and non-parametric bootstraps for estimating the MSE appear to have appealing properties regarding bias and stability.

Table 2.3: Performance of bootstrap MSE estimators in model-based simulation: mean and median of RB-RMSE and RRMSE-RMSE over areas.

| | *Normal-Poisson* | | *Interaction-Poisson* | | *NB3* | | *NB1* | |
|---|---|---|---|---|---|---|---|---|
| | Median | Mean | Median | Mean | Median | Mean | Median | Mean |
| RB-RMSE[%] | | | | | | | | |
| GMERF P | 3.57 | 3.45 | −0.36 | 5.33 | | | | |
| GMERF SP | 8.23 | 8.21 | −2.24 | −1.29 | 5.81 | 6.71 | −8.54 | 2.21 |
| MERF NPC | 1.97 | 2.03 | 0.28 | 0.48 | 3.95 | 4.21 | −0.83 | −1.21 |
| RRMSE-RMSE[%] | | | | | | | | |
| GMERF P | 10.16 | 10.56 | 47.91 | 57.39 | | | | |
| GMERF SP | 23.01 | 26.26 | 24.42 | 24.77 | 42.79 | 47.21 | 60.75 | 74.41 |
| MERF NPC | 7.90 | 8.36 | 12.86 | 13.29 | 15.27 | 15.67 | 23.97 | 24.35 |

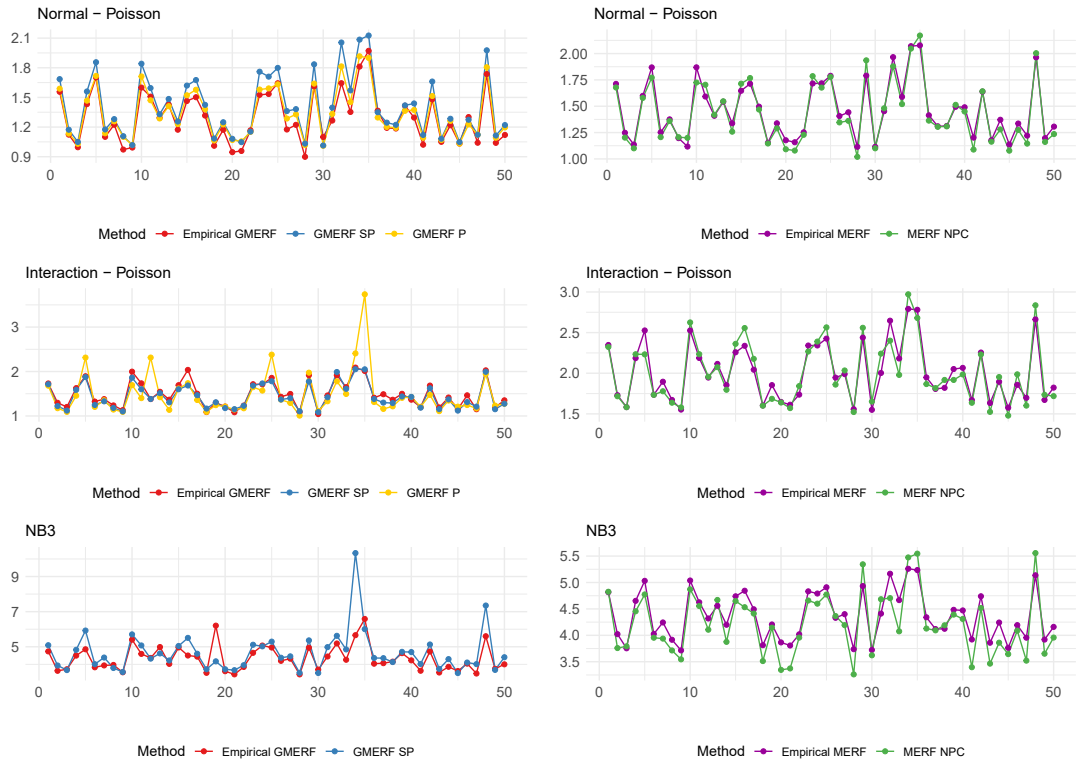



Figure 2.2: Empirical and bootstrapped area-level RMSEs for three scenarios.

## 2.5 Application to education data from Guerrero-Mexico

In this section, we describe the application of our SAE methods to estimate the average number of school years completed by women in the municipalities of Guerrero, Mexico, based on data from 2010.

We combine two primary data sources: the Mexican household income and expenditure survey ENIGH (Encuesta Nacional de Ingreso y Gastos de los Hogares) and census microdata provided by the National Institute of Statistics and Geography (Instituto Nacional de Estadística y Geografía). The ENIGH survey encompasses data from $1,445$ households spread across $40$ municipalities in Guerrero, while the census dataset includes information on $112,265$ households from all $81$ municipalities in the state. The survey samples are notably varied, with individual municipality samples ranging from a minimum of $8$ households to a maximum of $458$, and a median of $21$ households per municipality. This disparity in sample sizes results in $41$ municipalities being left out of the survey sample, highlighting a significant challenge in applying SAE methods effectively. Table 2.4 summarizes these domain-specific data characteristics, providing an overview of the coverage and distribution of data points across the municipalities.

The implementation of SAE methodologies involves the development of a working model using survey data, which is subsequently refined with census or administrative data. In the context of Poisson GLMMs, such as the EBPP approach, selecting the appropriate variables is crucial. We apply the Akaike information criterion to identify the optimal model for predicting the average number of school years completed by women. This process resulted in the selection

Table 2.4: Summary statistics on in- and out-of-sample areas: area-specific sample size of census and survey data.

| Municipalities | Total 81 | | In-sample 40 | | Out-of-sample 41 | |
|---|---|---|---|---|---|---|
| | Min. | Q1 | Median | Mean | Q3 | Max. |
| Survey area sizes | 8.00 | 15.00 | 21.00 | 36.12 | 32.00 | 458.00 |
| Census area sizes | 361.00 | 619.00 | 833.00 | 1,386.00 | 1,656.00 | 6,297.00 |

of 21 out of 39 potential covariates for our final Poisson GLMM model.

In contrast to the explicit variable selection required for Poisson GLMMs, random forests implement an implicit model selection process (Breiman, 2001). Figure 2.3 illustrates the insights gained from this method through partial dependence plots (PDP) and variable importance plots (VIP) (Greenwell, 2017; Greenwell et al., 2020). The PDP quantifies the marginal effect of specific predictors on the target variable, highlighting the non-linear dynamics between predictors and outcomes. The VIP ranks the significance of each predictor based on the mean decrease in impurity (variance), which is aggregated from the number of splits across all trees that involve the predictor. Notably, key predictors such as years of working experience (`jexp`), household income (`inglabpc`) and the level of education in relation to average (`escol_rel_hog`) emerge as critical for modeling the educational attainment of women. Overall, the analysis in Figure 2.3 underscores the complex interplay between predictors and the target variable, affirming the value of advanced modeling techniques in uncovering these relationships in the context of SAE.

Figure 2.4 presents an analysis of Pearson residuals from the Poisson GLMM applied to the estimation of school years for women in Guerrero municipalities. The histogram displays a positive skew in the residuals and the presence of some large values. This pattern is further evidenced by the plot showing the distribution of residuals across municipalities, with several exhibiting numerous positive residuals. The red dashed lines in the right plot of Figure 2.4 indicate the values -2 and 2, beyond which the Pearson residuals may suggest a poor model fit. Several residuals can be seen exceeding these values. These observations suggest a potential issue with overdispersion - a critical aspect considering the equidispersion assumption of Poisson models. To verify this, we conducted overdispersion tests comparing Poisson and Negative Binomial models. Both the Likelihood Ratio Test (LRT) and Dean's PB test indicate significant overdispersion, with LRT yielding a statistic of $67.471$ and $p < 2.2e^{-16}$, and Dean's PB test showing $6.840$ with $p = 3.953e^{-12}$.

Additionally, Figure 2.5 illustrates a plot of raw residuals against fitted values, revealing a distinct pattern across the range of fitted values. In particular, for predicted school years between 0 and 10, the residuals exhibit higher variability and a discernible concentration of negative values, indicating potential difficulties in accurately predicting outcomes for individuals with fewer years of education. This pattern may indicate the presence of unaccounted variability or non-linear effects within this range, which could be indicative of potential misspecification in the Poisson GLMM. Given these findings, we argue that the model's limitations may warrant the exploration of alternative, more flexible modeling approaches. Consequently,

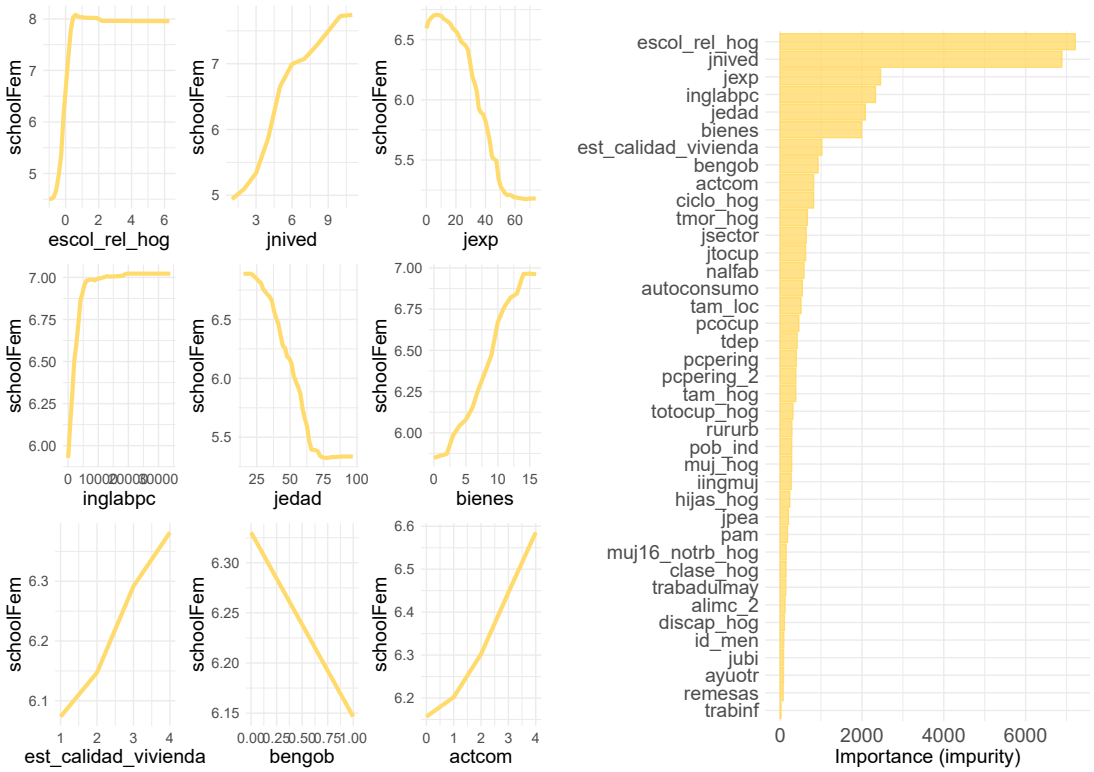



Figure 2.3: Visual diagnostics on predictive relations between dependent variable *schoolFem* and predictors. Partial dependency plots (left-hand side); variable importance plot (right-hand side).

we estimate both the GMERF and the MERF. These data-driven semi-parametric models could potentially offer an improved fit and robustness over the traditional parametric EBPP, which is based on a Poisson GLMM.

Figure 2.6 showcases the results from three different estimation methods applied. These methods include direct estimation, which is feasible for 40 of the 81 municipalities, as well as the model-based approaches provided by GMERF and MERF. The direct estimates serve as a baseline, feasible only for municipalities with sufficient data. Conversely, GMERF and MERF extend our insights into regions where direct data are lacking, thereby enhancing our understanding of regional disparities in educational attainment. All three methods illustrate clear regional differences in educational outcomes among the municipalities. However, the point estimates from GMERF and MERF are similar, suggesting that both model-based methods provide consistent estimates despite their methodological differences. Furthermore, the maps provide useful information on the geographical distribution of the average number of years of schooling for women.

Figure 2.7 highlights the variability in the estimation accuracy for the average number of school years for women, as measured by the coefficients of variation (CV). We use the calibrated bootstrap method from the R package **emdi** (Kreutzmann et al., 2019) for direct estimates and the semi- and non-parametric bootstrap procedures in Section 2.3 and Appendix B.1 for GMERF and MERF, each involving $B = 200$ bootstrap replications. Given the established presence of overdispersion, we opted to forgo the use of the parametric bootstrap in this appli-

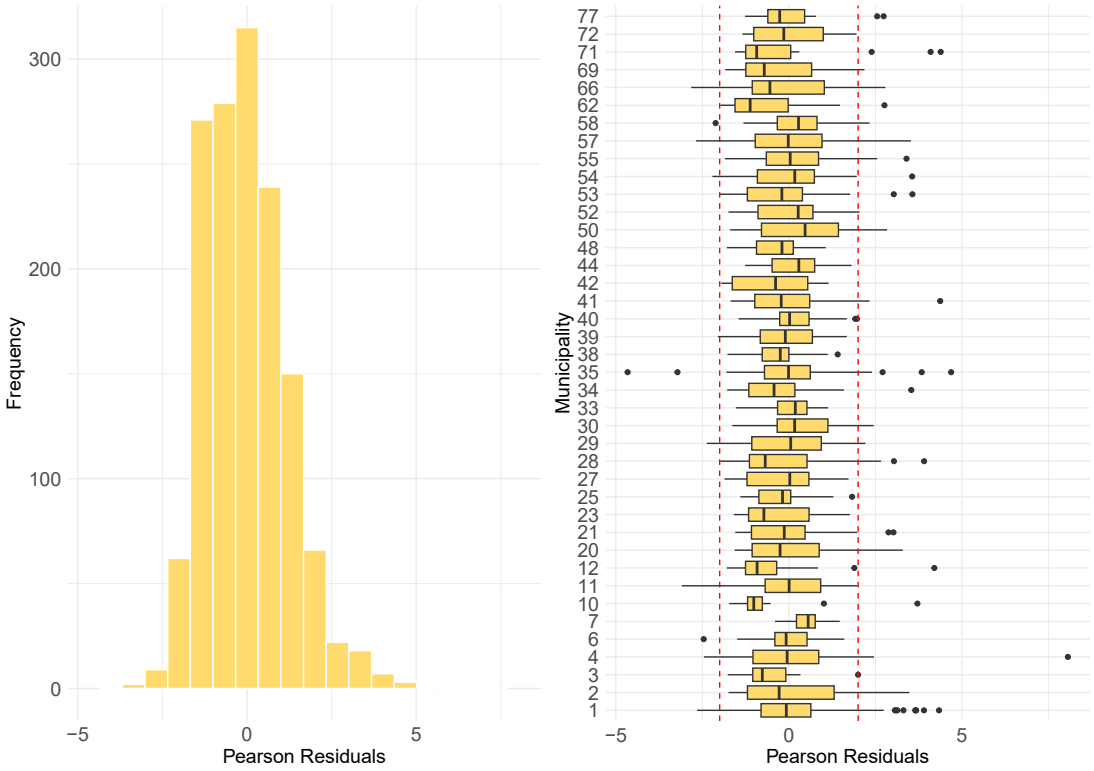



Figure 2.4: Model fit diagnostics for the Poisson GLMM: Histogram of Pearson residuals (left) and box-plots of Pearson residuals by municipalities (right).

cation. For both in-sample and out-of-sample domains, the comparison between GMERF and MERF reveals no notable differences in precision, suggesting comparable performance across these model-based methods.

The observed similarities in CVs across different machine learning models and bootstrap methods could be attributed to the underlying strength of overdispersion in the data, as evidenced by a dispersion ratio of 1.50 in the estimated Poisson GLMM. This corresponds closely with the scale parameter of 3.071 in the Negative Binomial distribution used in the analysis, which aligns well with the model-based simulation results for the *NB3* setting (scale parameter of 3) focusing solely on in-sample domains. It seems that there are no substantial differences between GMERF and MERF in a data scenario that shows only slight to moderate overdispersion.

Further validation of these findings is provided in Section 2.6, where a design-based simulation study enhances the reliability of the point and MSE estimators discussed here, allowing for a more detailed exploration of their performance.

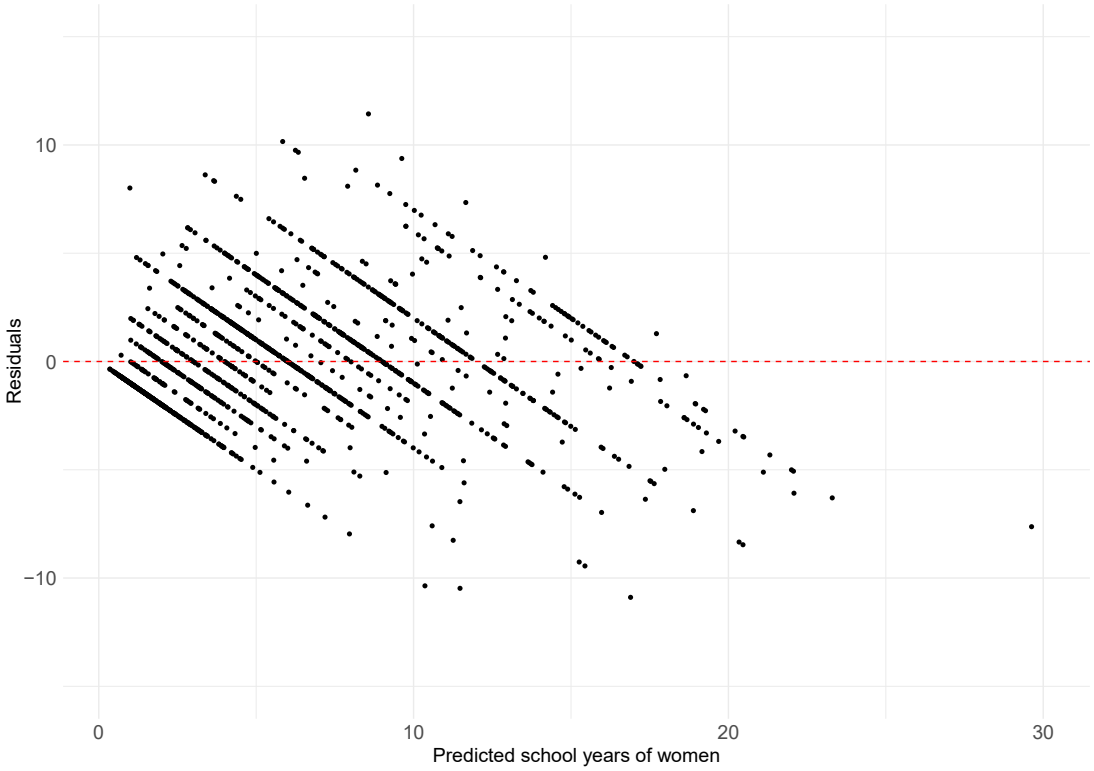



Figure 2.5: Model fit diagnostics for the Poisson GLMM: Raw residuals vs. predicted values.

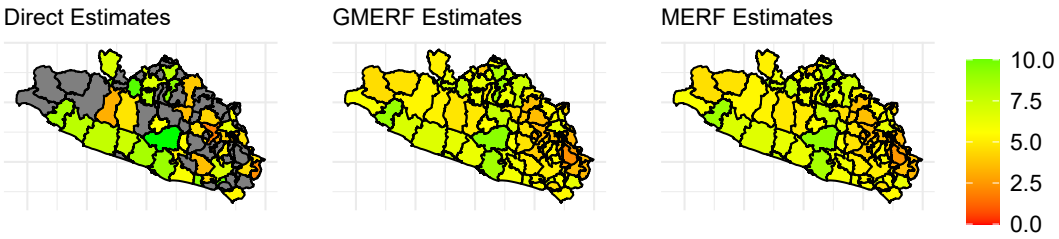



Figure 2.6: Estimated school years of women for the state of Guerrero based on direct estimates, GMERF and MERF.

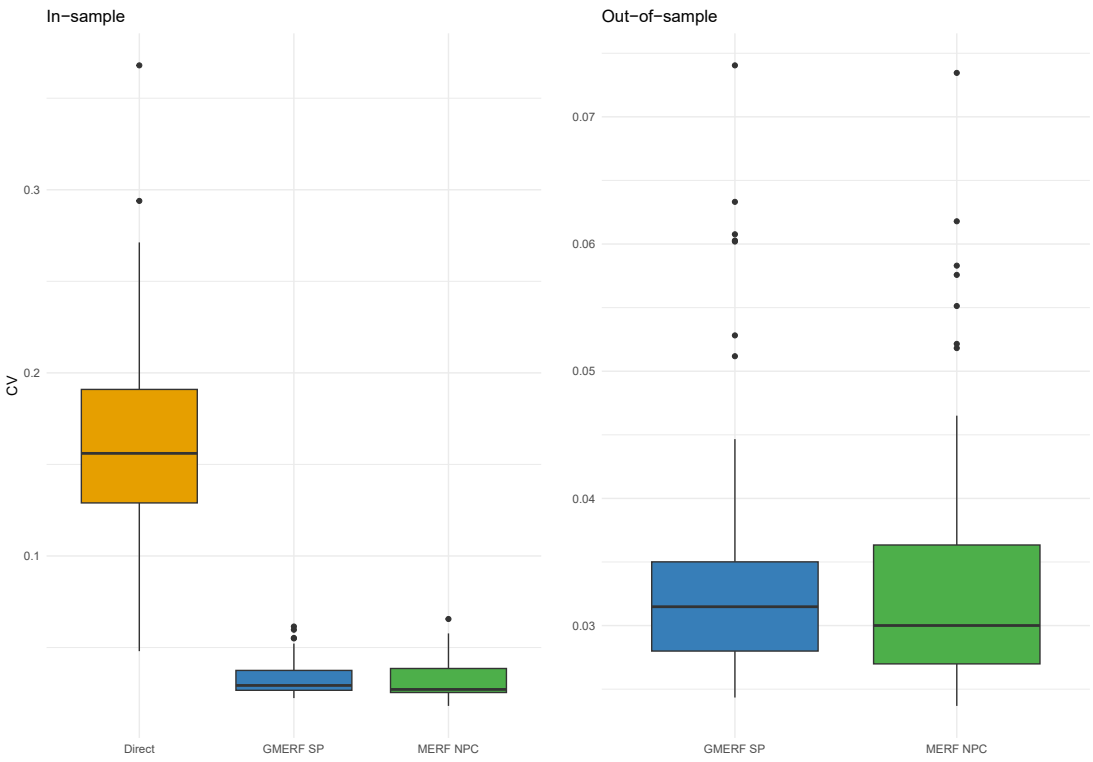



Figure 2.7: Area-specific CVs for Means for in- and out-of-sample areas.

## 2.6 Design-based simulation

This section assesses the efficacy of the proposed small area estimation methods by employing a design-based simulation, using the same dataset and structure as the real-data application discussed in Section 2.5. The simulation was conducted with $500$ independent pseudo-survey samples drawn from the fixed population census dataset of Guerrero. Each pseudo-sample mirrors the characteristics of the original survey, maintaining the same number of in-sample municipalities. This ensures that each of the $500$ pseudo-survey samples has consistent structure and overall sample size. The true values are defined as the area-level means derived from the original census data. The estimation methods evaluated - EBPP, GMERF, and MERF - are applied as described in Section 5, using the same fixed working model for EBPP throughout the simulation.

Figures 2.8 and 2.9, as well as Table 2.5, present the results in terms of bias and RMSE for the estimated area-level means in Guerrero. A direct comparison between GMERF and MERF reveals no substantial difference in RMSE, indicating similar performance levels for both methods across different sample domains. Notably, both GMERF and MERF demonstrate superior performance in terms of RMSE when compared to EBPP, particularly highlighting the strength of non-parametric approaches in this simulation context.

Table 2.6 evaluates the performance of the proposed MSE bootstrap procedures in terms of relative bias of RMSE and relative RMSE of RMSE. In terms of RRMSE-RMSE, the adjusted non-parametric bootstrap method for the MERF appears to be the most efficient method in this design-based simulation. Regardless of the bootstrap and estimation method used, the RB-RMSE for the in-sample areas indicates an acceptable bias regarding the median and mean RB-RMSE. For out-of-sample areas, we encounter underestimation in terms of the median value and overestimation according to mean values for all three methods.

Table 2.5: Mean and median for Bias and RMSE over total, in- and out-of-sample areas for point estimates.

| Municipalities | Total | | In-sample | | Out-of-sample | |
|---|---|---|---|---|---|---|
| | Median | Mean | Median | Mean | Median | Mean |
| Bias | | | | | | |
| EBPP | $0.1257$ | $0.0067$ | $0.1549$ | $0.1356$ | $-0.1168$ | $-0.1192$ |
| GMERF | $-0.0096$ | $0.0110$ | $0.0367$ | $0.0311$ | $-0.0693$ | $-0.0085$ |
| MERF | $0.0349$ | $0.0631$ | $0.0686$ | $0.0750$ | $-0.0190$ | $0.0515$ |
| RMSE | | | | | | |
| EBPP | $0.4174$ | $0.4422$ | $0.4563$ | $0.4711$ | $0.3510$ | $0.4139$ |
| GMERF | $0.2209$ | $0.2747$ | $0.1933$ | $0.2462$ | $0.2666$ | $0.3918$ |
| MERF | $0.2152$ | $0.2670$ | $0.2051$ | $0.2411$ | $0.2460$ | $0.2923$ |

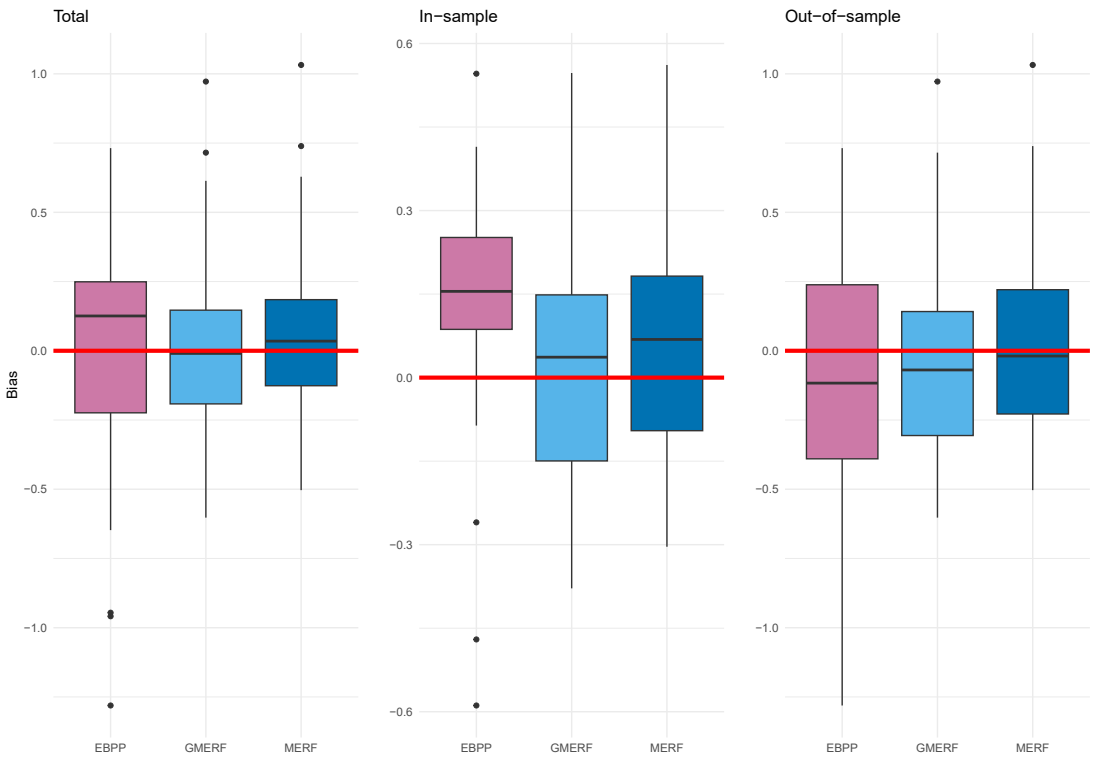



Figure 2.8: Bias of area-specific point estimates including details on in- and out-of-sample areas. Comparison of empirical bias from the design-based simulation for target variable *schoolfem*.

Table 2.6: Mean and median for relative Bias and relative RMSE of the estimated RMSE over total, in- and out-of-sample areas for MSE estimates.

| Municipalities | Total | | In-sample | | Out-of-sample | |
|---|---|---|---|---|---|---|
| | Median | Mean | Median | Mean | Median | Mean |
| RB-RMSE[%] | | | | | | |
| GMERF P | −5.85 | 2.17 | −2.95 | 0.27 | −16.50 | 4.03 |
| GMERF SP | −4.16 | 8.14 | 3.21 | 7.19 | −12.10 | 9.07 |
| MERF NPC | 6.44 | 17.59 | 8.08 | 13.17 | −4.38 | 21.90 |
| RRMSE-RMSE[%] | | | | | | |
| GMERF P | 53.63 | 57.97 | 45.14 | 48.23 | 58.93 | 67.47 |
| GMERF SP | 51.07 | 58.38 | 46.89 | 50.61 | 55.79 | 65.95 |
| MERF NPC | 42.68 | 56.11 | 40.84 | 46.07 | 45.09 | 65.90 |

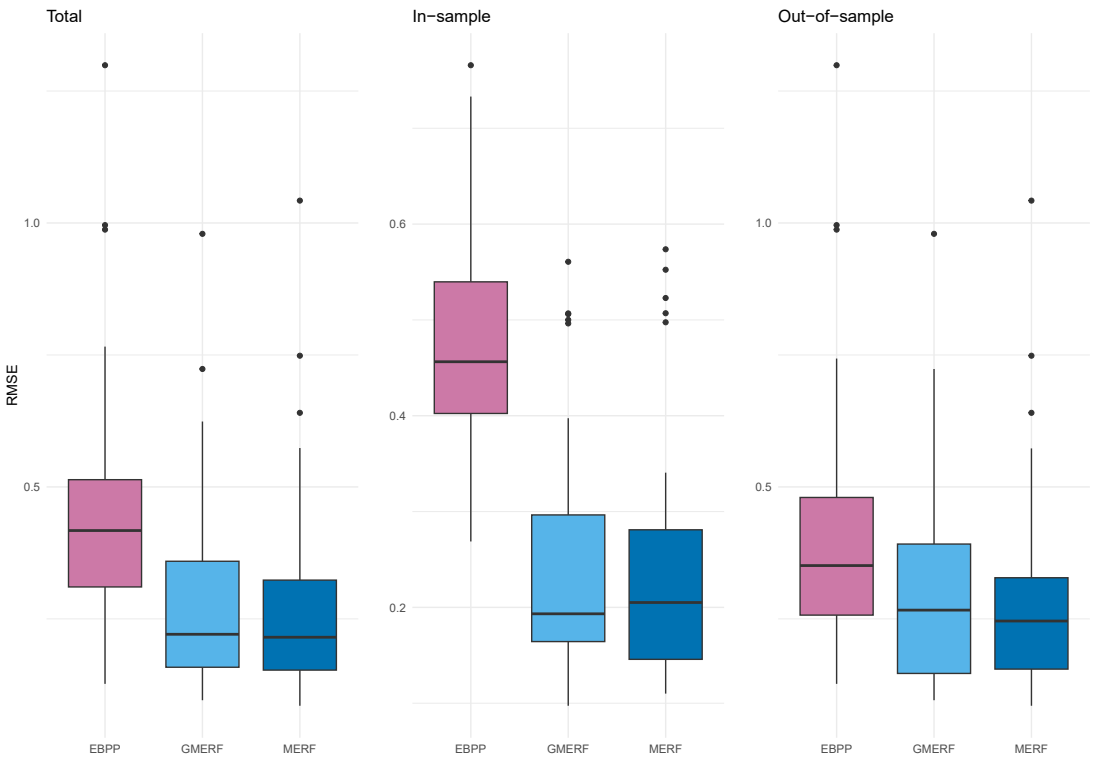



Figure 2.9: Performance of area-specific point estimates including details on in- and out-of-sample areas. Comparison of empirical RMSEs from the design-based simulation for target variable *schoolfem*.

## 2.7 Conclusion

In this paper, we introduce generalized tree-based machine learning methods for estimating disaggregated, small area means for count data. Additionally, we explore the impact of overdispersion on the performance of these estimation methods. Our findings reveal that the MERF, which operates independently of distributional assumptions with respect to the count data, performs superior under conditions of severe overdispersion compared to the GMERF. Conversely, the GMERF shows better performance when the assumptions of the Poisson distribution are met or only moderately violated.

In Section 2.2, we extend the GMERF procedure (Frink and Schmid, 2024a) to accommodate count data and discuss generalized semi-parametric mixed models at the unit-level, treating Poisson GLMM-based SAE methods such as EBPP as special cases. Furthermore, the MERF approach (Krennmair and Schmid, 2022), initially developed for continuous data, is adapted for discrete outcome variables, exploring potential advantages and limitations, especially in relation to overdispersion. We present also three bootstrap methods for evaluating point estimators for area-level means: a parametric and a semi-parametric MSE bootstrap procedure for the GMERF, and an adjusted non-parametric MSE bootstrap procedure for the MERF (originally for continuous outcomes). The efficacy of point and MSE estimates is assessed through model-based simulations. Additionally, Section 2.5 applies these methods to data from the Mexican state of Guerrero.

The simulations in Section 2.4 demonstrate that the GMERF and MERF point estimates surpass traditional methods when dealing with non-linear interactions between covariates at the linear predictor level. Notably, in scenarios of overdispersion, the MERF remains strong performance without dependence on Poisson distribution assumptions, outperforming generalized models. The MSE-bootstrap schemes proposed are validated as reliable based on their performance in our simulations. The practical application in Guerrero and further design-based simulations in our Appendix also support the effectiveness of these approaches.

From a methodological standpoint, the scope for further research into the application of GMERF in SAE is broad and promising. One potential direction is to adapt GMERF to model non-linear indicators, such as quantiles. This could involve applying the smearing technique introduced by Chambers and Dunstan (1986) and further developed for machine-learning methods by Krennmair et al. (2023). Furthermore, extending the GMERF to accommodate the Negative Binomial distribution by incorporating an additional dispersion parameter into the model could address issues of overdispersion more effectively than models based solely on the Poisson distribution. This modification would potentially allow GMERF to provide more accurate estimates for count data characterized by high variability, making it a more versatile tool in the field of SAE. These expansions not only enhance the flexibility and applicability of GMERF but also open up new possibilities for addressing complex problems in statistical estimation. Finally, further research is needed to explore the theoretical properties of the discussed bootstrap schemes for area-level means. Existing theoretical results for random forests, such as the work of Wager et al. (2014) and Wager and Athey (2018) on uncertainty quantification and the consistency of individual (unit-level) predictions, could provide a foundation for developing theoretical results for MERF- and GMERF-related bootstrap schemes for area-level means.

## Declarations

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

## Acknowledgments

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. The authors are grateful for the computation time provided by the HPC service of the Freie Universität Berlin. Finally, the authors are indebted to the Co-Editor, Associate Editor and three referees for comments that significantly improved the paper.

# Appendix B

## B.1 Non-parametric bootstrap for MERF with count data

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

2. Utilizing the marginal residuals $z_{ij}$, determine 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. Derive the vector of level-1 residuals using $\hat{z}_{ij} = z_{ij} - \bar{z}_i$. Following Krennmair and Schmid (2022), the residuals are adjusted to the bias-corrected variance and centered, indicated by $\hat{z}^c_{ij}$. Level-2 residuals $\bar{z}_i$ are additionally adjusted for the estimated variance $\hat{\sigma}^2_\nu$ and centered, indicated by $\bar{z}^c$.

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

   (a) Independently draw samples with replacement from the scaled and centered 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) Compute $\eta^{(b)}_{ij} = \hat{f}(\mathbf{x}_{ij}) + \bar{z}^{(b)}_i + z^{(b)}_{ij}$.

   (c) To obtain a count target variable, we match $\eta^{(b)}_{ij}$ with the set of estimated unit-level predictors from the sample $\left\{\hat{\eta}_t |\ \hat{\eta}_t = \hat{f}(\mathbf{x}_{it}) + \hat{\nu}_i,\ \text{for}\ t = 1, ..., n\right\}$ by finding the corresponding index $\tilde{t}$ solving

$$\min_t \left|\eta^{(b)}_{ij} - \hat{\eta}_t\right|.$$

   (d) Define the bootstrap population by $y^{(b)}_{ij} = y_{\tilde{t}}$ for $j = 1, \ldots, N$ and calculate the true bootstrap population mean $\mu^{(b)}_i$ for $i = 1, ..., D$.

   (e) For each bootstrap population, select a bootstrap sample that matches the original sample size $n_i$ and estimate $\hat{f}^{(b)}()$ and $\hat{\nu}^{(b)}_i$. Obtain estimates for the mean $\hat{\mu}^{(b)}_i$.

5. Utilizing the $B$ bootstrap samples, the MSE estimator is computed as follows:

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

# Chapter 3

# A framework for generalizing mixed effect random forests in R

**Abstract**

The R package **SAEforest** simplifies the estimation of regionally disaggregated indicators using machine learning techniques for small area estimation. It provides tools for model presentation and diagnostics. The package version 1.0.0 includes mixed effect random forests for continuous outcomes. Since version 2.0.0, the package has incorporated generalized mixed effect random forests for binary and count-based indicators. To assess the uncertainty of the area-level estimates, corresponding mean squared error estimators are implemented. Additionally, version 2.0.0 introduces two new diagnostic plots and an updated hyperparameter tuning function for the generalized random forest components. The functionality of these enhancements is illustrated with examples using synthetic datasets for Austrian districts.

**Keywords**: Generalized linear mixed models, Machine learning, Small area estimation, Survey statistics

## 3.1 Introduction

Small area estimation (SAE) plays a crucial role not only in academic research but also in various applied fields, offering deeper insights into local-level indicators. It is particularly beneficial for those producing official statistics and policymakers who rely on the accurate estimation of disaggregated indicators. Traditional surveys often fail to provide data at fine levels of granularity, typically due to small sample sizes that undermine the precision of estimates. SAE techniques provide a cost-effective solution to this problem, avoiding the need for expensive and time-consuming expansions of survey sample sizes (Ghosh, 2020).

The essence of SAE lies in amalgamating disparate data sources through model-based frameworks. By supplementing existing survey data with auxiliary information, the accuracy of indicator estimation at the area-level is considerably enhanced. Areas or domains can refer to either geographical regions or specific subpopulations within the wider population of interest (Rao and Molina, 2015; Tzavidis et al., 2018). To simplify the application of SAE methods for users, various R (R Core Team, 2024) packages are available, including **rsae** (Schoch, 2012), **JoSAE** (Breidenbach, 2015), **sae** (Molina and Marhuenda, 2015), **rhnerm** (Sugasawa, 2016), **saeRobust** (Warnholz, 2018), **emdi** (Kreutzmann et al., 2019) and **saeTrafo** (Würz, 2024).

In SAE, mixed models serve as a cornerstone in the analysis of unit-level survey data and the estimation of area-level indicators. These models adeptly address the hierarchical nature of observations by incorporating random effects. One instantiation of this model is the nested error regression model (Battese et al., 1988), which necessitates access to unit-level survey information and administrative auxiliary data. Nevertheless, it is imperative to acknowledge the constraints posed by the distributional and structural assumptions inherent in these models, as they may not consistently align with the realities encountered in SAE scenarios (Rao and Molina, 2015). One potential solution to the limitations of mixed models is the use of machine learning techniques. These methods extend beyond parametric models by learning predictive relationships from data, including higher-order interactions between covariates, without the need for explicit model assumptions (Varian, 2014). Within the broad spectrum of machine learning techniques, the **SAEforest** package (Krennmair, 2022) emphasizes tree-based models, specifically random forests (Breiman, 2001), due to their outstanding predictive performance in the presence of outliers and their capability to handle model selection challenges (Biau and Scornet, 2016).

While machine learning methods have been applied in SAE (Blumenstock et al., 2015; Bilton et al., 2017; Pokhriyal and Jacques, 2017; Steele et al., 2017; Bilton et al., 2020; Hersh et al., 2021), they have drawbacks such as neglecting subpopulation correlations and struggling with complex covariance structures. Integrating tree-based methods into mixed effects models has gained traction to overcome these limitations. For instance, mixed effect regression trees (MERT) (Hajjem et al., 2011) replace the linear combination of covariates in linear mixed models (LMM) fixed effects with a regression tree (Breiman et al., 1984). To enhance prediction accuracy, researchers have delved into the integration of random forests within mixed effects models (Hajjem et al., 2014). Contemporary R packages for continuous outcomes, dependent data, and tree-based machine learning methods typically overlook SAE topics, often prioritizing prediction over inference. Examples include **MixRF** (Wang and Chen, 2016), **splinetree** (Neufeld and Heggeseth, 2019), **LongituRF** (Capitaine, 2020) and **RandomforestGLS** (Saha et al., 2021). Krennmair and Schmid (2022) introduce the concept of mixed effect random forest (MERF) for continuous outcomes within the methodological framework of SAE, while Krennmair (2022) presents the initial version 1.0.0 of the **SAEforest** package.

Recent advancements have broadened the applicability of methods originally designed for continuous response variables to various types of responses. For instance, Hajjem et al. (2017) introduce generalized mixed effect regression trees (GMERT), extending MERT to discrete data. Fontana et al. (2021) introduce generalized mixed effects tree (GMET) for classification tasks, using tree leaves as indicator variables rather than using the tree predictions employed in GMERT. To incorporate tree ensembles, Pellagatti et al. (2021) extend the GMET approach by using random forests rather than standard trees in the fixed effects component of the mixed effects model. The R package **glmertree** (Fokkema and Zeileis, 2023) is the only package which deals with tree-based methods for dependent data with discrete target variables. However, it does not concentrate on SAE applications. Ultimately, Frink and Schmid (2024a,b) extend the GMERT approach to create a generalized mixed effect random forest (GMERF), integrating it into the SAE framework for estimating area-level means for both binary and count outcomes.

Therefore, the latest update of the **SAEforest** package, version 2.0.0, brings several enhancements to the existing version:

- Introducing the new workflow of the updated `SAEforest_model` function, which facilitates the utilization of different machine learning-based mixed models.
- Providing point estimates, as well as their parametric and non-parametric uncertainty estimators for the GMERF.
- Offering an adjusted non-parametric uncertainty estimator specifically tailored for the MERF (originally designed for continuous target variables) with discrete outcomes.
- Introducing an `aggregate_to` argument, akin to the **emdi** package, enabling users to specify the desired target domain-level for displaying the results, independent of the level at which the random effect is specified.
- Adding two supplementary diagnostic plots to the `plot` function to aid in model evaluation.
- Implementing an updated hyperparameter tuning function, `SAEforest_tuning`, designed for optimizing the components of the random forest for the MERF and GMERF models.

The structure of the paper is as follows: Section 3.2 introduces the new statistical methodology implemented in version 2.0.0 of the **SAEforest** package. Section 3.3 describes the example datasets included in the package. Section 3.4 presents the core functionalities of **SAEforest**, providing a general overview of the updated main function `SAEforest_model` and presenting generic methods for the GMERF model. Section 3.5 summarizes methods and results, also outlining potential further extensions.

## 3.2 Statistical methods

MERF models have been available since the initial release of the package, version 1.0.0 (Krennmair, 2022). These encompass the estimation of means using unit-level covariates (Krennmair and Schmid, 2022), the estimation of means leveraging aggregated covariate information (Krennmair et al., 2022), and the estimation of non-linear indicators employing unit-level covariates (Krennmair et al., 2023). To enhance the usability of machine learning-style mixed models for discrete target variables, version 2.0.0 of the **SAEforest** package incorporates GMERFs for binary and count outcomes, including their uncertainty estimates (Frink and Schmid, 2024a,b). However, the GMERF estimator depends on the assumptions of the discrete distribution functions. Conversely, the MERF, originally designed for continuous indicators, may offer a robust alternative for modeling discrete outcomes, as it is not dependent on these distributional assumptions (Frink and Schmid, 2024b). Consequently, version 2.0.0 also integrates MERFs specifically for discrete target variables, along with an adjusted uncertainty estimator.

### 3.2.1 Generalized machine learning-type unit-level mixed model

Assume a finite population, denoted by $P$, which is comprised of $D$ disjoint domains $P_1, P_2, \ldots, P_D$, each of which is represented by a sub-population of size $N_i$, where $i = 1, ..., D$. The discrete unit-level outcome variable $y_{ij}$, for individual $j$ in domain $i$ is available for every unit within the sample. The sample $s$ is composed of domain-specific sub-samples $s_i$ with an overall size of $n = \sum_1^D n_i$. Conversely, non-sampled observations are characterized as $r_i$ with a size of $N_i - n_i$. Individual units within each area are labeled as $j \in s_i$ for sampled observations and $j \in r_i$ for unsampled observations. The total population size of all domains is defined as $N = \sum_{i=1}^D N_i$. The vector $\mathbf{x}_{ij} = [x_1, x_2, ..., x_p]^\intercal$ includes $p$ covariates, and the explanatory variables are available for every unit in the population. We assume that $y_{ij}$ follows a unit-level machine learning-type mixed model:

$$E(y_{ij}|\nu_i) = \mu_{ij}, \quad \mu_{ij} = g(\eta_{ij}), \quad \eta_{ij} = f(\mathbf{x}_{ij}) + \nu_i, \quad \nu_i \sim N(0, \sigma_\nu^2). \tag{3.1}$$

The fixed part, $f()$, represents a machine learning method and expresses the conditional mean of the linear predictor ($\eta_{ij}$) given the covariates. In the R package **SAEforest**, the function $f()$ is the random forest. The hierarchical structure of observations is represented by area-specific random intercepts $\nu_i$ with variance $\sigma_\nu^2$. The term $g()$ denotes a differentiable, monotonic link function that defines the relation between the mean ($\mu_{ij}$) and the systematic component. Version 2.0.0 of the R package offers four different distributions and link functions: `gaussian`, `binomial`, `poisson` and `quasipoisson`. If $y_{ij}$ follows a Gaussian distribution, the link function simplifies to the identity function. As a result, in model (3.1), the GMERF transforms into the MERF, as suggested by Krennmair and Schmid (2022), and is operationalized through the `MERFranger` function within the **SAEforest** package.

Computationally, GMERFs rely on a doubly iterative process, with micro iterations nested within macro iterations to obtain optimal estimates on model components $\hat{f}()$, $\hat{\nu}_i$ and $\hat{\sigma}_\nu^2$. During each macro iteration, updates are made to the initial estimates of a linearized target variable ($y_{L,ij}$) and to the weights ($w_{ij}$):

$$y_{L,ij} = g(\mu_{ij}) + (y_{ij} - \mu_{ij})g'(\mu_{ij}) \quad \text{and} \quad w_{ij} = (v_{ij}g'(\mu_{ij})^2)^{-1},$$

with $v_{ij}$ representing a known variance function. The updated values of the linearized response variable and weights are used as the response variable and weights for the following micro iterations, respectively. The procedure employs micro iterations that are similar to the Expectation Maximization (EM) algorithm (Moon, 1996). This method is used iteratively to determine the parameters of model (3.1). The process involves two main steps: first, estimating the forest function with the assumption that the random effects term is accurate, using the **ranger** package (Wright and Ziegler, 2017); second, estimating the random effects component, assuming the Out-of-Bag predictions (OOB-predictions) from the forest are accurate, using the **lme4** package (Bates et al., 2015). To elucidate, a weighted MERF pseudo-model is estimated in the second step:

$$y_{L,ij} = \hat{f}(\mathbf{x}_{ij})^{\text{OOB}} + \hat{\nu}_i + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$ are individual independently distributed unit-level errors, which are mutually independent to the random effects. Convergence of the micro iterations is tracked through marginal alterations in the log-likelihood of the model, while the convergence of the macro iterations is assessed by monitoring changes in the linear predictor. Computationally, the GMERF algorithm is implemented in the function `GMERFranger` of **SAEforest**. For additional methodological insights, refer to Frink and Schmid (2024a,b).

### 3.2.2 Generalized small area averages

The GMERF estimator for individual predictions can be expressed as:

$$\begin{aligned}\hat{\mu}_{ij}^{\text{GMERF}} &= g\left(\hat{f}(\mathbf{x}_{ij}) + \hat{\nu}_i\right) \\ &= g\left(\hat{f}(\mathbf{x}_{ij}) + \frac{\hat{\sigma}^2_\nu}{\hat{\sigma}^2_\nu + \hat{\sigma}^2_\epsilon / n_i}\left(\frac{1}{n_i}\sum_{j \in s_i}(y_{L,ij} - \hat{f}(\mathbf{x}_{ij})^{\text{OOB}})\right)\right).\end{aligned} \tag{3.2}$$

However, using the **SAEforest** package, our focus is on constructing area-level estimators. Therefore, depending on the chosen link function, the proposed estimators for the area-level mean are as follows:

$$\begin{aligned}\hat{\mu}_{i,\texttt{gaussian}}^{\text{GMERF}} &= \hat{\mu}_i^{\text{MERF}} \\ &= \frac{1}{N_i}\sum_{j \in P_i}\hat{f}(\mathbf{x}_{ij}) + \frac{\hat{\sigma}^2_\nu}{\hat{\sigma}^2_\nu + \hat{\sigma}^2_\epsilon / n_i}\left(\frac{1}{n_i}\sum_{j \in s_i}(y_{L,ij} - \hat{f}(\mathbf{x}_{ij})^{\text{OOB}})\right) \\ &= \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \hat{\nu}_i,\end{aligned} \tag{3.3}$$

$$\hat{\mu}_{i,\texttt{binomial}}^{\text{GMERF}} = \frac{\exp\left(\bar{\hat{f}}_i(\mathbf{x}_{ij}) + \hat{\nu}_i\right)}{1 + \exp\left(\bar{\hat{f}}_i(\mathbf{x}_{ij}) + \hat{\nu}_i\right)}, \tag{3.4}$$

$$\begin{aligned}\hat{\mu}_{i,\texttt{poisson}}^{\text{GMERF}} &= \hat{\mu}_{i,\texttt{quasipoisson}}^{\text{GMERF}} \\ &= \exp\left(\bar{\hat{f}}_i(\mathbf{x}_{ij}) + \hat{\nu}_i\right).\end{aligned} \tag{3.5}$$

For out-of-sample areas, the proposed estimators simplify to the fixed component:

$$\hat{\mu}_{i,\texttt{gaussian}}^{\text{GMERF}} = \hat{\mu}_i^{\text{MERF}} = \bar{\hat{f}}_i(\mathbf{x}_{ij}),$$

$$\hat{\mu}_{i,\texttt{binomial}}^{\text{GMERF}} = \frac{\exp\left(\bar{\hat{f}}_i(\mathbf{x}_{ij})\right)}{1 + \exp\left(\bar{\hat{f}}_i(\mathbf{x}_{ij})\right)},$$

$$\hat{\mu}^{\text{GMERF}}_{i,\texttt{poisson}} = \hat{\mu}^{\text{GMERF}}_{i,\texttt{quasipoisson}} = \exp\left(\bar{\hat{f}}_i(\mathbf{x}_{ij})\right).$$

### 3.2.3 Mean squared error estimation

To assess the precision of the estimates, the mean squared error (MSE) is the most frequently used measure in SAE (Rao and Molina, 2015). Version 1.0.0 of the **SAEforest** package provides a variety of MSE estimators. Version 2.0.0 introduces three additional bootstrap schemes - one parametric and two non-parametric - for estimating the MSE using machine learning-type mixed models of discrete outcomes. These bootstrap procedures primarily differ in the mechanism utilized for generating the population.

The first approach generates bootstrap realizations of the random effects parametrically and directly utilizes the distribution to create the target variable (`parametric`). This method is effective in scenarios where $y_{ij}$ follows a Binomial or Poisson distribution, as it explicitly leverages these specific distributions. Although distributions and link functions are commonly used in generalized models to depict the mean behavior of discrete data, the assumptions underlying these distributions are frequently not met. For example, with the Poisson distribution, overdispersion can occur, making the parametric bootstrap procedure inappropriate. To ensure robust uncertainty estimators, even in instances where the distribution's underlying assumptions are not met, we implement a non-parametric bootstrap approach for the GMERF (`nonparametricGmerf`). Additionally, following Frink and Schmid (2024b), we implement the adjusted non-parametric MSE bootstrap procedure for the MERF with discrete outcomes (`wild_discrete`) into the **SAEforest** package. Table 3.1 provides an overview of the included MSE approaches. For detailed formulas and derivations, please refer to the cited references.

Table 3.1: Overview of the MSE estimation techniques in the **SAEforest** package.

| Model | Target variable | Type of MSE | Reference |
|---|---|---|---|
| *MERF* | | | |
| `nonparametric` | Continuous | Non-parametric bootstrap | Krennmair and Schmid (2022) |
| `wild` | Continuous | Non-parametric bootstrap | Krennmair et al. (2023) |
| `wild_discrete` | Discrete | Non-parametric bootstrap | Frink and Schmid (2024b) |
| *GMERF* | | | |
| `parametric` | Discrete | Parametric bootstrap | Frink and Schmid (2024a,b) |
| `nonparametricGmerf` | Discrete | Non-parametric bootstrap | Frink and Schmid (2024b) |

## 3.3 Datasets for illustration

SAE primarily combines survey data with auxiliary information to increase the accuracy of the estimated indicator of interest. This additional information may encompass auxiliary covariates of domain-specific individual observations or domain-level aggregates. However, in

comparison to the MERF, the GMERF is currently only suitable for unit-level information. Consequently, estimation with aggregated covariates is not further elaborated upon.

The GMERF approach in the **SAEforest** package uses unit-level survey (`eusilcA_smp`) and unit-level auxiliary (`eusilcA_pop`) data from the **emdi** package (Kreutzmann et al., 2019). The authors provide a comprehensive account of the data generation process for the `eusilcP` dataset, which originates from the **simFrame** package (Alfons and Templ, 2013). This dataset comprises household-level synthetic data derived from the Austrian European Union Statistics on Income and Living Conditions (EU-SILC) from 2006. In the context of the **emdi** package, a spatially finer regional disaggregation was manually created by employing a random assignment method, with the objective of considering the diverse regional income levels across Austria. The synthetic population dataset thus encompasses the $94$ Austrian districts as the lowest regional level. The population data comprises of $25,000$ households, while the unit-level sample data, created using stratified random sampling, includes $1,945$ households. The sample data covers $70$ districts, with $24$ areas out-of-sample.

Alongside domain-level identifiers for states (`state`) and districts (`districts`), auxiliary variables include $14$ socio-demographic characteristics such as gender or receipt of state benefits. Both datasets include equivalized household income (`eqIncome`) as a potential target variable, which is continuous rather than discrete. Consequently, a binary outcome variable (`Ybin`) and a count outcome variable (`Ycount`) have been constructed from this continuous variable for the practical examples of GMERFs in Section 3.4. For further detailed information, please refer to `help(eusilcA_smp)` and `help(eusilcA_pop)`.

## 3.4 Functionalities and practical examples

The structure of this section is outlined as follows: Section 3.4.1 provides an overview of the updated primary function, `SAEforest_model`, delineating its functionality and the workflow in R. In Sections 3.4.2 and 3.4.3, we demonstrate the practical application of the `SAEforest_model` function for binary and count outcomes, respectively, using exemplary data. Furthermore, Section 3.4.4 highlights the versatility of the generic methods provided by **SAEforest** for the analysis and visualization of the associated S3 object resulting from the GMERF algorithm. Finally, in Section 3.4.5, we present the updated tuning function for the hyperparameters of GMERF and MERF models.

### 3.4.1 Overview `SAEforest_model`

The `SAEforest_model` function provides the methodologies outlined in Section 3.2 and presented in Krennmair (2022), enabling the invocation of both the GMERF and the MERF approaches through this function. Table 3.2 provides an overview of the $28$ input arguments of `SAEforest_model`, designed to accommodate various distributions and link functions, thereby supporting a wide array of specifications. Each argument is accompanied by a brief description, default settings (if specified), and information regarding its availability for GMERF or MERF. It is important to note that not every argument needs to be specified for every estimated model. At a minimum, the input requires the target variable (`Y`), predictive covariates

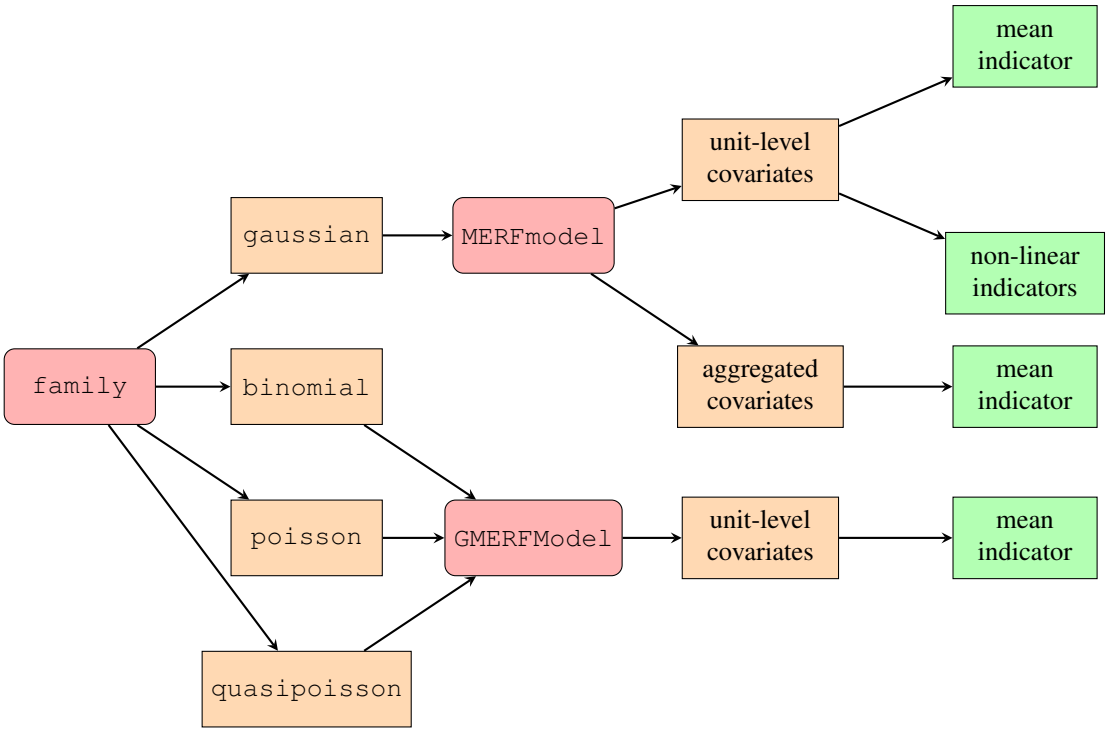



Figure 3.1: Overview of the estimation process.

(`X`), domain identifier (`dName`), sample data (`smp_data`), and population data (`pop_data`).

The flow diagram depicted in Figure 3.1 illustrates the available estimation options. The choice of the distribution function specified by the `family` argument determines whether a MERF or GMERF model is estimated. If `family = "gaussian"`, a `MERFmodel` is estimated following the methodology outlined by Krennmair and Schmid (2022), Krennmair et al. (2022) and Krennmair et al. (2023), incorporating both unit-level and aggregated covariates. Conversely, when `family` is set to `"binomial"`, `"poisson"`, or `"quasipoisson"`, a `GMERFmodel` is estimated following the approaches described by Frink and Schmid (2024a,b), where only unit-level covariates can be used. `SAEforest_model` is the wrapper function for the core functionalities of both `GMERFranger` and `MERFranger`. For more comprehensive details, please consult the `help(GMERFranger)` function.

The output of the `SAEforest_model` function is an object of class `SAEforest`, comprising no fewer than three elements:

1. `GMERFmodel` or `MERFmodel` (depending on the chosen estimation strategy).
2. Point estimates (`Indicators`).
3. `MSE_estimates` if requested (otherwise `NULL`).

Table C.1 offers a comprehensive summary of the elements of `SAEforest` objects and provides a detailed explanation of the components of a `GMERFmodel`.

To evaluate the resulting `GMERFmodel` or `MERFmodel` objects, users can utilize the generic functions `summary` and `plot` to examine summary statistics and visual model diagnostics, respectively. These generic methods operate on S3 objects of class `SAEforest`

Table 3.2: Input arguments of main function `SAEforest_model`.

| Arguments | Short description | Default | Available for | |
|---|---|---|---|---|
| | | | MERF | GMERF |
| `Y` | Target variable | | ✓ | ✓ |
| `X` | Covariates | | ✓ | ✓ |
| `family` | Distribution and link function | `"gaussian"` | ✓ | ✓ |
| `dName` | Domain identifier | | ✓ | ✓ |
| `smp_data` | Survey data | | ✓ | ✓ |
| `pop_data` | Census or administrative data | | ✓ | ✓ |
| `MSE` | MSE estimation | `"none"` | ✓ | ✓ |
| `importance` | Variable importance mode used by the random forest | `"impurity"` | ✓ | ✓ |
| `B` | Number of bootstrap replications for MSE estimation | 100 | ✓ | ✓ |
| `initialRandomEffects` | Initial estimate of random effects | 0 | ✓ | ✓ |
| `ErrorToleranceEm` | Value to monitor the EM algorithm's convergence | 0.0001 | ✓ | ✓ |
| `MaxIterationsEm` | Maximal amount of iterations for the EM algorithm | 25 | ✓ | ✓ |
| `ErrorTolerancePql` | Value to monitor the PQL algorithm's convergence | 0.001 | | ✓ |
| `MaxIterationsPql` | Maximal amount of iterations for the PQL algorithm | 10 | | ✓ |
| `...` | Additional parameters passed on to `ranger` | | ✓ | ✓ |
| `na.rm` | Remove missing values | `TRUE` | ✓ | ✓ |
| `meanOnly` | Calculating domain-level means only | `TRUE` | ✓ | |
| `B_adj` | Number of bootstrap replications for adjustment of residual variance | 100 | ✓ | ✓ |
| `aggData` | Use aggregated auxiliary information | `FALSE` | ✓ | |
| `popnsize` | Information of population size of domains | `NULL` | ✓ | |
| `OOsample_obs` | Out-of-sample observations from the closest area | 25 | ✓ | |
| `ADDsamp_obs` | Out-of-sample observations from the closest area if first iteration fails | 0 | ✓ | |
| `w_min` | Minimal number of covariates from which informative weights are computed | 3 | ✓ | |
| `threshold` | Custom threshold | `NULL` | ✓ | |
| `custom_indicator` | Additional functions containing the indicators to be computed | `NULL` | ✓ | |
| `smearing` | Smearing approach or a MC-based version for point estimates | `TRUE` | ✓ | |
| `B_MC` | Number of bootstrap populations to be generated for the MC version | 100 | ✓ | |
| `aggregate_to` | Area-level for which results are displayed | `NULL` | ✓ | |

(Chambers and Hastie, 1992). Additionally, the function `summarize_indicators` extracts the final area-specific estimates, while `SAEforest_tuning` assesses potential improvements by tuning the model's hyperparameters. Lastly, the function `map_indicators` can be employed to visualize the estimators. Detailed examples of the functionality of these methods are provided in the subsequent subsections.

### 3.4.2 Estimation procedure for the GMERF

In this practical example, we make use of the synthetic Austrian EU-SILC dataset mentioned in Section 3.3. The goal is to illustrate the estimation of means for small areas using the GMERF approach for binary and count outcomes, specifically at the level of the 94 Austrian districts. The necessary sample and population data are readily available within the **SAEforest** package for convenient access.

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

R> # Create binary input variable and specify the explanatory
R> # variables
R> Ybin <- ifelse(eusilcA_smp$eqIncome
+                 >0.6*median(eusilcA_smp$eqIncome),0,1)
R> X <- eusilcA_smp[,-c(1, 16, 17, 18)]
```

In order to prepare the covariates (`X`) for analysis, only predictor variables should be included. Consequently, we exclude variables that contain domain-level codes and the response variable. It is crucial to explicitly specify the domain names (`dName`) to differentiate between different domains for random intercepts. The survey dataset, referred to as `smp_data`, and the dataset comprising auxiliary information, referred to as `pop_data`, are designated accordingly. In the following example, the objective is to estimate a binary indicator (`Ybin`). Thus, we need to set the `family` parameter to `"binomial"`. To perform the parametric bootstrap with 50 repetitions and obtain uncertainty measures, we set `MSE = "parametric"` and `B = 50`.

```
R> fit1 <- SAEforest_model(Y = Ybin,
+                           X = X,
+                           family = "binomial",
+                           dName = "district",
+                           smp_data = eusilcA_smp,
+                           pop_data = eusilcA_pop,
+                           MSE = "parametric",
+                           B = 50)
```

In addition to binary outcomes, the GMERF can also be used to estimate count data following either a Poisson or Quasi-Poisson distribution. In the subsequent brief example, the

`SAEForest_model` function is employed for count data with `family = "poisson"` and applies a non-parametric bootstrap method with `MSE = "nonparametricGmerf"`. Furthermore, the function enables users to directly provide parameters to the `ranger` function using the three-dotted option `(...)`. Key parameters to specify in a (generalized) random forest include the number of trees (`num.trees`) and the number of variables to randomize for each node split decision (`mtry`) (Krennmair, 2022).

```
R> # Create exemplary discrete target variable
R> Ycount <- as.numeric(cut(eusilcA_smp$eqIncome,
+                           breaks = 7,
+                           labels = 1:7))

R> fit2 <- SAEforest_model(Y = Ycount,
+                          X = X,
+                          family = "poisson",
+                          dName = "district",
+                          smp_data = eusilcA_smp,
+                          pop_data = eusilcA_pop,
+                          MSE = "nonparametricGmerf",
+                          B = 50,
+                          mtry = 3,
+                          num.trees = 500)
```

### 3.4.3 Estimation procedure for the MERF with discrete outcomes

Furthermore, the updated `SAEforest_model` function now offers the option of using the MERF to estimate not only continuous outcomes but also binary and count responses. For all three options, the `family` parameter is set to `"gaussian"`. One advantage of the MERF over the GMERF is that it is not based on the distribution assumptions of the Binomial and Poisson distributions. Consequently, it has the potential to lead to more accurate results if the aforementioned assumptions are violated (Frink and Schmid, 2024b).

```
R> # Binary outcome
R> fit3 <- SAEforest_model(Y = Ybin,
+                          X = X,
+                          family = "gaussian",
+                          dName = "district",
+                          smp_data = eusilcA_smp,
+                          pop_data = eusilcA_pop)
```

In addition to the point estimate, the adjusted non-parametric MSE estimator for the MERF, specified as `MSE = "wild_discrete"`, can also be employed:

```
R> # Count outcome
R> fit4 <- SAEforest_model(Y = Ycount,
```

```
+                              X = X,
+                              family = "gaussian",
+                              dName = "district",
+                              smp_data = eusilcA_smp,
+                              pop_data = eusilcA_pop,
+                              MSE = "wild_discrete",
+                              B = 50)
```

### 3.4.4 Generic methods

The main generic methods of the R package **SAEforest** (such as summary output, diagnostic plots, and graphic representation of estimates) are elaborated in detail, while all other functionalities are only concisely presented.

**Summary of `SAEforest` object**

An essential generic function, the `summary` function is available to obtain essential information and initial diagnostic results for a fitted `SAEforest` model object. It should be noted that there are slight differences in the output between a `MERFmodel` and a `GMERFmodel`. For instance, the latter displays the results of a weighted mixed model and provides information about the convergence of the GMERF algorithm from the `GMERFranger` function. Nevertheless, for both models, the `summary` function offers valuable insights into various SAE characteristics, including the number of out-of-sample and in-sample domains, the total number of observations, and the sample sizes specific to each area. Below is an illustrative example of the output obtained from applying the `summary` function to a fitted `GMERFmodel`:

```
R> summary(fit2)
```

```
_____________________________________________________________________
Generalized Mixed Effects Random Forest for Small Area Estimation
_____________________________________________________________________
Call:
SAEforest_model(Y = Ycount, X = X, family = "poisson",
dName = "district", smp_data = eusilcA_smp, pop_data = eusilcA_pop,
MSE = "nonparametricGmerf", B = 50, mtry = 3, num.trees = 500)

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):                      0.52221

Structural component of random effects:
```

---

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: res ~ -1 + (1 | district)
Data: data
Weights: pql_weigths

AIC      BIC   logLik   deviance  df.resid
296.9    308.1   -146.5    292.9     1943

Scaled residuals:
Min      1Q   Median      3Q     Max
-2.7550 -0.6446 -0.0284  0.5989  8.5768

Random effects:
Groups   Name        Variance Std.Dev.
district (Intercept) 0.03286  0.1813
Residual             0.10830  0.3291
Number of obs: 1945, groups:  district, 70

ICC:  0.2327625

Convergence of GMERF algorithm:
```

---

```
PQL convergence achieved after 7 iterations.
A maximum of 10 PQL iterations is used.
EM convergence achieved after 25 iterations.
A maximum of 25 EM iterations is used.
```

In this example, the output reveals that the synthetic Austrian dataset comprises 70 in-sample domains and 24 out-of-sample domains. It is noteworthy that this scenario is typical in SAE, where sample sizes specific to each area are relatively small, with a median of 22.5 observations. Regarding the random forest component, the `summary` output provides information on the tuning parameters used and the $R^2$. With an $R^2$ of 0.52, the model demonstrates a satisfactory level of predictive capability. The output also includes details about the weighted

mixed effects model employed in the GMERF algorithm. This information encompasses the variance of individual residuals, the variance of domain-level intercepts, and the intra-class correlation coefficient (ICC). The ICC indicates that approximately $23\%$ of the model's variance can be attributed to random effects. Lastly, the output highlights the convergence properties of the GMERF algorithm, offering reassurance regarding its reliability and stability.

**Diagnostic plots for (generalized) mixed effect random forests**

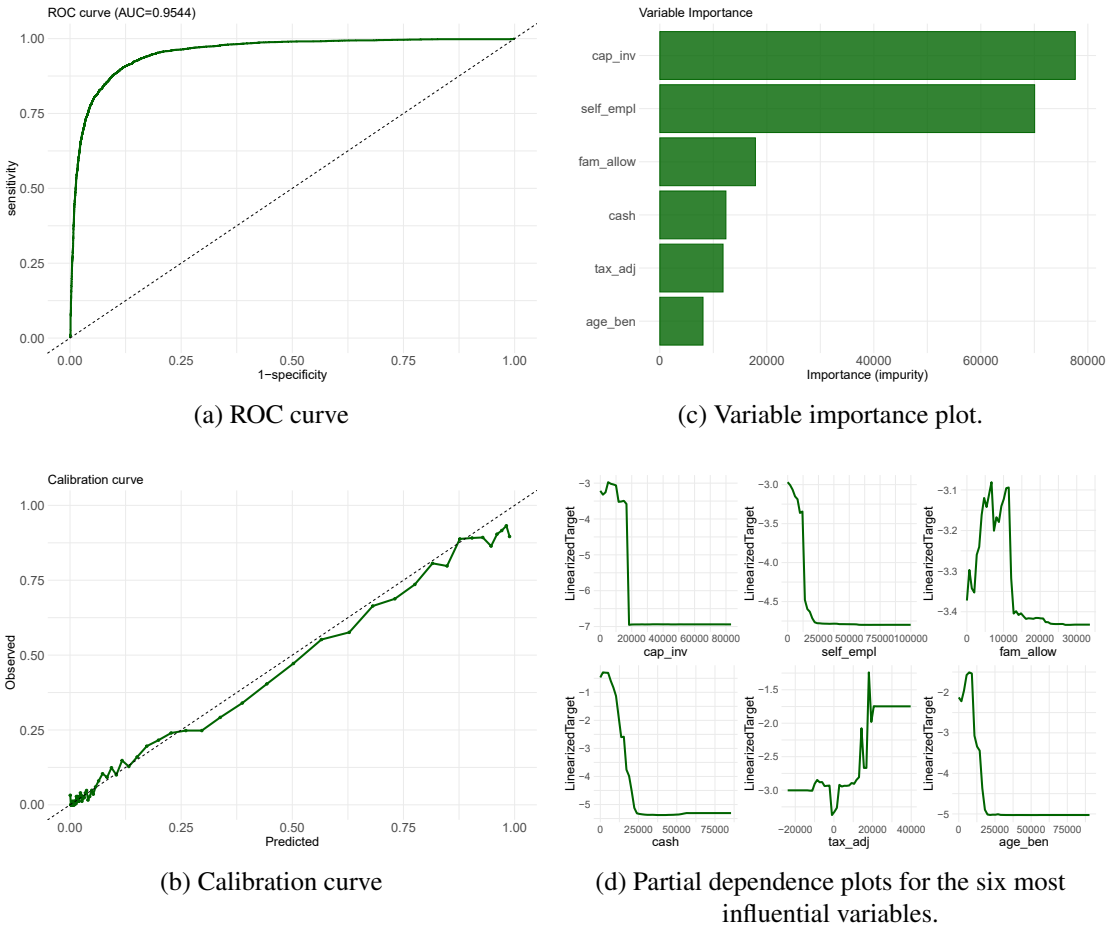
The **SAEforest** package provides a useful `plot` function that offers four diagnostic plots for analysis if `family = "binomial"`. Otherwise, only two plots are available. If a `GMERFmodel` is estimated and the population dataset also includes the target variable, users can specify the `add_data = TRUE` and `pop_data = eusilcA_pop` arguments to create plots that assess the effectiveness of GMERF in classifying observations into the 'true' category. These plots include the receiver operating characteristic (ROC) curve and calibration curve.

The ROC curve depicts the relationship between the false positive rate (1 - specificity) and true positive rate (sensitivity). Meanwhile, the area under the curve (AUC) quantifies the model's proficiency in distinguishing between the two classes. The ROC curve is generated using the **pROC** package (Robin et al., 2011). Furthermore, the calibration curve measures the model's accuracy in predicting probabilities by comparing the predicted values to the true values at various quantiles. A diagonal line on the calibration curve signifies a perfect correspondence between the predictions and the actual outcomes.

The other two plots are specific to random forest analysis and are also applicable if the `family` argument is not set to `"binomial"`. The first is the variable importance plot, which determines the importance of each variable based on the mean increase in individual mean squared prediction error when excluding that variable. The variable importance plot is generated using the `vip` function from the **vip** package (Greenwell et al., 2020). The final plot is the partial dependence plot, which displays the estimated marginal effect of a specific variable and illustrates whether the relationships are monotonic or more complex. The **pdp** package is employed to generate the partial dependence plots (Greenwell, 2017). Partial dependence plots are only output for numeric explanatory variables. If character or factor variables are present in the dataset, a warning message is issued.

```
R> # Create a binary outcome in the auxiliary dataset in order
R> # to generate the ROC and calibration curves
R> eusilcA_pop$Ybin <-ifelse(eusilcA_pop$eqIncome
+                       >0.6*median(eusilcA_pop$eqIncome),0,1)

R> # Display diagnostic plots
R> library(ggplot2)
R> library(pROC)
R> library(vip)
R> library(pdp)
R> plot(fit1, add_data = TRUE, pop_data = eusilcA_pop)
```




(a) ROC curve

(c) Variable importance plot.

(b) Calibration curve

(d) Partial dependence plots for the six most influential variables.

Figure 3.2: Output from function `plot`.

In the Austrian example, we obtain an AUC value of 0.9544 (Figure 3.2a). This high AUC signifies good classification and prediction accuracy, with values close to 1 indicating better separation between classes and improved prediction performance. Upon examination of the calibration curves (Figure 3.2b), the GMERF algorithm demonstrates great performance in predicting probabilities. The curves closely follow the diagonal line, indicating a strong correspondence between the predicted and true probabilities.

In the variable importance plots (Figure 3.2c), we identify the most important variables in the estimation process of the fixed effects. These include profits from capital investment (`cap_inv`), self-employment status (`self_empl`), family-related allowances (`fam_allow`), net cash income (`cash`), net repayments for tax adjustment (`tax_adj`) and age-related benefits (`age_ben`). These variables exert a considerable influence on the estimation results.

The partial dependence plots in Figure 3.2d reveal potential non-linear relationships, especially for variables such as `cap_inv`, `tax_adj` and `fam_allow`. These plots offer insights into the shape and direction of the relationships between these variables and the target variable, highlighting potential complexities beyond simple linear associations.

**Presentation of indicators**

The `summarize_indicators` function offers a convenient method to obtain point estimates, MSE, and coefficient of variation (CV) estimates from a `SAEforest` object. Ad-

ditionally, users can apply commonly known functions such as `tail`, `as.data.frame`, `as.matrix`, `head`, and `subset` to an S3 object created with this function, allowing for further analysis.

Below is a summary of the mean indicators and their corresponding CVs, with the option `CV = TRUE` set. This summary provides valuable information on the accuracy of the estimates.

```
R> head(summarize_indicators(fit2,
+                            MSE = FALSE,
+                            CV = TRUE))

          district     Mean    Mean_CV
           Amstetten 1.472638 0.04556439
               Baden 1.994440 0.03849310
             Bludenz 1.265336 0.06991515
      Braunau am Inn 1.361846 0.05286579
             Bregenz 2.785266 0.03844136
Bruck an der Leitha 2.017183 0.04594696
```

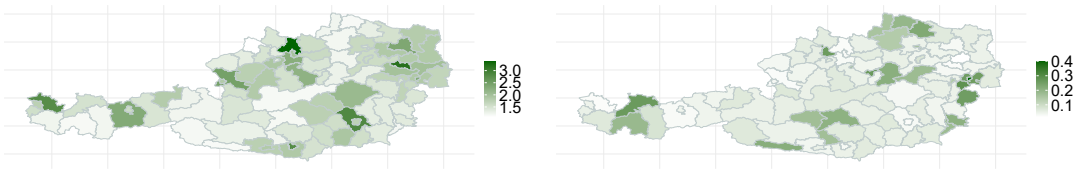
**Mapping of the area estimates**

The `map_indicators` function is employed to generate visualizations of estimates obtained from a fitted GMERF model object of class `SAEforest`. This function utilizes polygon data representing Austrian districts, which is conveniently available within the package. The `load_shapeaustria` function from **emdi** loads this map.

This function directly provides settings for the graphical representation, provides the processed data, and facilitates personalization of the map using **ggplot2** (Wickham, 2016). In case the area identifications within the `SpatialPolygonsDataFrame` and the S3 object do not match, the `map_tab` allows for the input of a dataset to assign the area identification. However, this is not the case in the following example. For more information, see `help(map_indicators)`.

```
R> library(ggplot2)
R> load_shapeaustria()
R> map_indicators(object = fit2,
+                 MSE = FALSE,
+                 CV = TRUE,
+                 map_obj = shape_austria_dis,
+                 map_dom_id = "PB")
```

Figure 3.3 showcases maps that depict the estimates of means and their corresponding CVs for all 94 Austrian districts estimated using a `GMERFmodel`. These maps offer valuable insights into the spatial patterns and variations across the districts.




(a) Mean (b) CV

Figure 3.3: Maps of point estimates and CVs of the mean for 94 districts in Austria.

**Supplementary generic methods**

In addition to the generic methods already discussed, the **SAEforest** package provides additional functionality through other generic functions. The `print` function can be used to obtain key information about the fitted model. As the `GMERFmodel` integrates random forest and mixed effects models, users can access specific elements directly from the fitted model object and make use of the generic functions from the **ranger** (Wright and Ziegler, 2017) and **lme4** (Bates et al., 2015) packages. These elements are contained in the objects `ForestModel` and `EffectModel`, respectively.

For objects of class `merMod`, there are particularly useful generics for extracting various model components. The functions `fixef`, `ranef`, `VarCorr`, `sigma`, and `getData` can be directly applied. For instance, the `fixef` function retrieves the fixed effects, while the `ranef` function extracts the random effects from the model.

### 3.4.5 Hyperparameter tuning

The `SAEforest_tuning` function is the latest tool for hyperparameter tuning in the package, succeeding the former `tune_parameters` function from version 1.0.0. It is crucial for fine-tuning the (generalized) random forest parameters. Depending on the `family` argument provided, the parameters of either a `MERFmodel` or `GMERFmodel` are adjusted. In both cases this function allows for the fine-tuning of key parameters: the number of variables considered for splitting at each node (`mtry`), the total number of trees in the forest (`num.trees`), the splitting rule for the trees (`splitrule`), and the minimum size of nodes (`min.node.size`).

To utilize the `SAEforest_tuning` function, specific inputs are required, including the control parameters for the `train` function from the **caret** package (Kuhn, 2008). These control parameters are supplied via the `trainControl` function from **caret** and include the type of cross-validation method (`method`), the number of folds (`number`), and the repetitions (`repeats`). Additionally, users need to predefine a grid of parameter candidates using the `expand.grid` function, outlining the potential values for the parameters being tuned.

The primary metrics for model fine-tuning include cross-validated results such as RMSE (root mean squared error), MAE (mean absolute error), and conditional $R^2$. In the specific example from Austria with a binary outcome variable, the optimal parameter for the number of

split candidates at each node is found to be 3.

```
R> library(caret)
R> gmerfControl <- trainControl(method = "repeatedcv",
+                               number = 5,
+                               repeats = 1)

R> gmerfGrid <- expand.grid(num.trees = 50,
+                           mtry = c(3, 7, 9),
+                           min.node.size = 10,
+                           splitrule = "variance")

R> SAEforest_tuning(Y = Ybin,
+                   X = X,
+                   family = "binomial",
+                   data = eusilcA_smp,
+                   dName = "district",
+                   trControl = gmerfControl,
+                   tuneGrid = gmerfGrid)

1945 samples
15 predictor
No pre-processing
Resampling: Cross-Validated (5 fold, repeated 1 times)
Summary of sample sizes: 1556, 1556, 1556, 1556, 1556
Resampling results across tuning parameters:
mtry  RMSE       Rsquared   MAE
3     0.2556887  0.5074106  0.1264698
7     0.2693485  0.4880614  0.1152952
9     0.2894488  0.4480750  0.1251385

Tuning parameter 'num.trees' was held constant at a value of 50
Tuning parameter 'min.node.size' held constant at a 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 = 50, mtry = 3,
min.node.size = 10 and splitrule = variance
```

## 3.5 Conclusion

In this paper, we delineate the expansion of the **SAEforest** package to version 2.0.0, where it incorporates different distributions and link functions, yielding GMERF models for unit-level information. Specifically, the package supports Binomial, Poisson, Quasi-Poisson, and

Gaussian distributions. By extending beyond the estimation of continuous target variables, this updated version offers a more versatile solution, accommodating a broader range of modeling needs in small area estimation. Besides the area point estimates, different bootstrap procedures for the domain uncertainty estimation, both parametric and non-parametric, are offered to the user. Furthermore, we implement the MERF tailored for discrete target variables, which operates independently of distribution assumptions, along with its MSE bootstrap estimation procedure. Both MERF and GMERF models can be conveniently estimated using a single function, `SAEforest_model`. Moreover, we have integrated new user-friendly diagnostic plots and introduced a novel tuning function for the components of the (generalized) random forest model.

Future versions of the package aim to incorporate additional machine learning methodologies, including Support Vector Machines, Gradient Boosting, and Bayesian Additive Regression Trees. Furthermore, the GMERF is to be developed for non-linear indicators and the use of aggregated covariates, with the intention of implementing it within the R package. Exploring the implementation of GMERF for other distributions, such as the Negative Binomial distribution, would be a compelling avenue for future development.

# Appendix C

Table C.1: The `SAEforest` object components distinguished in MERF and GMERF.

| Name | Short description | Available for | |
|---|---|---|---|
| | | MERF | GMERF |
| `MERFmodel` | Information on the model fit details on the MERF algorithm and variance components | ✓ | |
| `GMERFmodel` | Information on the model fit details on the GMERF algorithm and variance components | | ✓ |
| `Indicators` | Domain-level identifiers and estimates | ✓ | ✓ |
| `MSE_Estimates` | Domain-level identifiers and uncertainty estimates | ✓ | ✓ |
| `NrCovar` | Includes a list of variable names of covariates used for the calculation of calibration weights if `aggData = TRUE` | ✓ | |
| Details on `MERFmodel` and `GMERFmodel` | | | |
| `Forest` | Random forest of class `ranger` | ✓ | ✓ |
| `EffectModel` | Model of random effects of class `merMod` | ✓ | ✓ |
| `RandomEffects` | Random intercepts from `EffectModel` | ✓ | ✓ |
| `RanEffSD` | Standard deviation of random intercepts | ✓ | ✓ |
| `ErrorSD` | Standard deviation of unit-level errors | ✓ | ✓ |
| `VarianceCovariance` | VarCorr matrix from `EffectModel` | ✓ | ✓ |
| `LogLik` | Log-likelihood values of the EM algorithm | ✓ | ✓ |
| `unit_mu` | Unit-level $\mu_{ij}$ | | ✓ |
| `unit_eta` | Unit-level values of linear predictor $\eta_{ij}$ | | ✓ |
| `weight` | Weights from the PQL-algorithm | | ✓ |
| `yLin` | Linearized response variable | | ✓ |
| `IterationsUsedPql` | Iterations used until convergence of the PQL algorithm | | ✓ |
| `MaxIterationsPql` | Maximal amount of iterations for the PQL algorithm | | ✓ |
| `IterationsUsedEm` | Iterations used until convergence of the EM algorithm | ✓ | ✓ |
| `MaxIterationsEm` | Maximal amount of iterations for the EM algorithm | ✓ | ✓ |
| `family` | Error distribution and link function | | ✓ |
| `OOBresiduals` | OOB residuals | ✓ | |
| `call` | Call for the object | ✓ | ✓ |
| `data_specs` | Data characteristics | ✓ | ✓ |
| `data` | Survey data | ✓ | ✓ |

# 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.

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

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

Battese, G., 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.

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.

Blumenstock, J., G. Cadamuro, and R. On (2015). Predicting poverty and wealth from mobile phone metadata. Science 350(6264), 1073–1076.

Boubeta, M., M. J. Lombardía, and D. Morales (2016). Empirical best prediction under area-level poisson mixed models. TEST 25(3), 548–569.

Boubeta, M., M. J. Lombardía, and D. Morales (2017). Poisson mixed models for studying the poverty in small areas. Computational Statistics & Data Analysis 107, 32–47.

Breidenbach, J. (2015). **JoSAE**: Functions for some unit-level small area estimators and their variances. R package version 0.2.3.

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

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

Breiman, L., J. Friedman, C. Stone, and R. Olshen (1984). Classification and regression trees. Taylor & Francis.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88(421), 9–25.

Browne, W. J., S. V. Subramanian, K. Jones, and H. Goldstein (2005). Variance partitioning in multilevel logistic models that exhibit overdispersion. Journal of the Royal Statistical Society: Series A (Statistics in Society) 168(3), 599–613.

Capitaine, L. (2020). **LongituRF**: Random forests for longitudinal data. R package version 0.9.

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., E. Dreassi, and N. Salvati (2014). Disease mapping via negative binomial regression M-quantiles. Statistics in Medicine 33(27), 4805–4824.

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

Chambers, R., N. Salvati, and N. Tzavidis (2016). Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. Journal of the Royal Statistical Society: Series A (Statistics in Society) 179(2), 453–479.

Chandra, H., S. Kumar, and K. Aditya (2018). Small area estimation of proportions with different levels of auxiliary data. Biometrical Journal 60(2), 395–415.

Chandra, H., N. Salvati, and R. Chambers (2017). Small area prediction of counts under a non-stationary spatial model. Spatial statistics 20, 30–56.

Chen, S., J. Jiang, and T. Nguyen (2015). Observed best prediction for small area counts. Journal of Survey Statistics and Methodology 3(2), 136–161.

Consejo Nacional de Evaluación de la Política de Desarrollo Social (2020). Informe de pobreza y evaluación. Technical report.

Datta, G. S. and M. Ghosh (2012). Small area shrinkage estimation. Statistical Science 27(1), 95–114.

Dreassi, E., M. G. Ranalli, and N. Salvati (2014). Semiparametric M-quantile regression for count data. Statistical Methods in Medical Research 23(6), 591–610.

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

Fahrmeir, L. and G. Tutz (2001). Models for multicategorical responses: Multivariate extensions of generalized linear models. In Multivariate Statistical Modelling Based on Generalized Linear Models, pp. 69–137. New York: Springer.

Farrell, P. J., B. MacGibbon, and T. J. Tomberlin (1997). Bootstrap adjustments for empirical bayes interval estimates of small-area proportions. The Canadian Journal of Statistics / La Revue Canadienne de Statistique 25(1), 75–89.

Flores-Agreda, D. and E. Cantoni (2019). Bootstrap estimation of uncertainty in prediction for generalized linear mixed models. Computational Statistics & Data Analysis 130, 1–17.

Fokkema, M. and A. Zeileis (2023). **glmertree**: Generalized linear mixed model trees. R package version 0.2.4.

Fontana, L., C. Masci, F. Ieva, and A. M. Paganoni (2021). Performing learning analytics via generalised mixed-effects trees. Data 6(7).

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

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29(5), 1189–1232.

Frink, N. and T. Schmid (2024a). Small area estimation with generalized random forests: Estimating poverty rates in Mexico. Preprint: `https://arxiv.org/abs/2406.03861`.

Frink, N. and T. Schmid (2024b). Small area prediction of counts under machine learning-type mixed models. Preprint: `https://arxiv.org/abs/2407.05849`.

Ghosh, M. (2020). Small area estimation: Its evolution in five decades. Statistics in Transition. New Series 21(4), 1–22.

Ghosh, M., K. Natarajan, T. W. F. Stroud, and B. P. Carlin (1998). Generalized linear models for small-area estimation. Journal of the American Statistical Association 93(441), 273–282.

Ghosh, M. and J. N. K. Rao (1994). Small area estimation: An appraisal. Statistical Science 9(1), 55–76.

Goldstein, H., W. Browne, and J. Rasbash (2002). Partitioning variation in multilevel models. Understanding Statistics 1(4), 223–231.

González-Manteiga, W., M. Lombardía, I. Molina, D. Morales, and L. Santamaría (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. Computational Statistics & Data Analysis 51(5), 2720–2733.

González-Manteiga, W., M. 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.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Applications to poisson models. Econometrica 52(3), 701–720.

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.

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.

Hajjem, A., F. Bellavance, and D. Larocque (2017). Generalized mixed effects regression trees. Statistics & Probability Letters 126, 114–118.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.

Hersh, J., R. Engstrom, and M. Mann (2021). Open data for algorithms: Mapping poverty in Belize using open satellite derived features and machine learning. Information Technology for Development 27(2), 263–292.

Hobza, T. and D. Morales (2016). Empirical best prediction under unit-level logit mixed models. Journal of Official Statistics 32(3), 661–692.

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

Instituto Nacional de Estadística y Geografía (2017). Anuario estadístico y geográfico de Tlaxcala 2017. Technical report.

Instituto Nacional de Estadística y Geografía (2021). Aspectos geográficos: Tlaxcala. Technical report.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. Journal of Statistical Planning and Inference 111(1), 117–127.

Jiang, J. and P. Lahiri (2001). Empirical best prediction for small area inference with binary data. Annals of the Institute of Statistical Mathematics 53(2), 1572–9052.

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

Jiang, J. and T. Nguyen (2021). Linear and generalized linear mixed models and their applications (2 ed.). 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.

Kern, C., T. Klausch, and F. Kreuter (2019). Tree-based machine learning methods for survey research. Survey Research Methods 13(1), 73–93.

Krennmair, P. (2022). The R package **SAEforest**. R package version 1.0.0.

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

Krennmair, P., T. Schmid, and N. Tzavidis (2023). The estimation of poverty indicators using mixed effects random forests: Case study for the Mexican state of Veracruz. Working 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.

Kuhn, M. (2008). Building predictive models in R using the **caret** package. Journal of Statistical Software 28(5), 1–26.

Lehtonen, R. and A. Veijanen (2009). Design-based methods of estimation for domains and small areas. In C. Rao (Ed.), Handbook of Statistics. Elsevier.

Long, J. S. (1997). Regression models for categorical and limited dependent variables: Advanced quantitative techniques in the social sciences, Volume 7. Sage Publications, Thousand Oaks.

MacGibbon, B. and T. J. Tomberlin (1989). Small area estimates of proportions via empirical bayes techniques. Survey Methodology 15, 237–252.

Malec, D., J. Sedransk, C. L. Moriarity, and F. B. Leclere (1997). Small area inference for binary variables in the national health interview survey. Journal of the American Statistical Association 92(439), 815–826.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. Journal of the American Statistical Association 92(437), 162–170.

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

Moon, T. K. (1996). The expectation-maximization algorithm. IEEE Signal Processing Magazine 13(6), 47–60.

Morales, D., M. D. Esteban, A. Pérez, and T. Hobza (2021). A course on small area estimation and mixed models. Springer Cham.

Nandram, B., J. Sedransk, and L. Pickle (1999). Bayesian analysis of mortality rates for U.S. health service areas. Sankhya: The Indian Journal of Statistics, Series B 61(1), 145–165.

National Center for Health Statistics (2023). National health interview survey.

Neufeld, A. and B. Heggeseth (2019). **splinetree**: Longitudinal regression trees and forests. R package version 0.2.0.

Organisation for Economic Co-operation and Development (2012). Income distribution data review - Mexico. Technical report.

Parker, P. A., R. Janicki, and S. H. Holan (2023). Comparison of unit-level small area estimation modeling approaches for survey data under informative sampling. Journal of Survey Statistics and Methodology 11(4), 858–872.

Pellagatti, M., C. Masci, F. Ieva, and A. M. Paganoni (2021). Generalized mixed-effects random forest: A flexible approach to predict university student dropout. Statistical Analysis and Data Mining: The ASA Data Science Journal 14(3), 241–257.

Peragine, V., M. G. Pittau, E. Savaglio, and S. Vannucci (2021). On multidimensional poverty rankings of binary attributes. Journal of Public Economic Theory 23(2), 248–274.

Pfeffermann, D. (2011). Small area estimation. In M. Lovric (Ed.), International Encyclopedia of Statistical Science, pp. 1346–1349. Springer Berlin Heidelberg.

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

Pokhriyal, N. and D. C. Jacques (2017). Combining disparate data sources for improved poverty prediction and mapping. Proceedings of the National Academy of Sciences 114(46), E9783–E9792.

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.

Rao, J. N. K. and I. Molina (2015). Small area estimation. Wiley, Hoboken, New Jersey.

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller (2011). **pROC**: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77.

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.

Runge, M. (2023). Estimating intra-regional inequality with an application to German spatial planning regions. Journal of Official Statistics 39(2), 203–228.

Saha, A., S. Basu, and A. Datta (2021). **RandomForestsGLS**: Random forests for dependent data. R package version 0.1.3.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science 2(3), 160.

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. (2012). Robust unit-level small area estimation: A fast algorithm for large datasets. Austrian Journal of Statistics 41(4), 243–265.

Statista (2021). Average annual wage in Mexico from 2000 to 2021. `https://www.statista.com/statistics/812354/mexico-average-annual-wage/`. accessed 2023-05-16.

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

Stroup, W. W. (2012). Generalized linear mixed models: Modern concepts, methods and applications. CRC Press.

Sugasawa, S. (2016). **rhnerm**: Random heteroscedastic nested error regression. R package version 1.1.

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

Tarozzi, A. and A. Deaton (2009). Using census and survey data to estimate poverty and inequality for small areas. The Review of Economics and Statistics 91(4), 773–792.

Tzavidis, N., M. G. Ranalli, N. Salvati, E. Dreassi, and R. Chambers (2015). Robust small area prediction for counts. Statistical Methods in Medical Research 24(3), 373–395.

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 (2015). Transforming our world: The 2030 agenda for sustainable development. Technical report.

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

Ver Hoef, J. M. and P. L. Boveng (2007). Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? Ecology 88(11), 2766–2772.

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 clustered incomplete data. R package version 1.0.

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.

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

World Bank Group (2020). COVID-19 to add as many as 150 million extreme poor by 2021. Technical report.

Wright, M. N. and A. Ziegler (2017). **ranger**: A fast implementation of random forests for highdimensional 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. John Wiley & Sons.

Würz, N. (2024). **saeTrafo**: Transformations for unit-level small area models. R package version 1.0.4.