Analysing Opportunity Cost of Care Work using Mixed Effects Random Forests under Aggregated Census Data

Patrick Krennmair[*], Nora Würz[*], and Timo Schmid[**]

[*]Institute of Statistics and Econometrics, Freie Universität Berlin, Berlin, Germany

[**]Department of Statistics and Econometrics, Otto-Friedrich-Universität Bamberg, Bamberg,

Germany

Abstract

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

Keywords: Official statistics; Small area estimation; Mean squared error; Tree-based methods

1 Introduction

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

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

In SAE, several strategies evolved to prevent model-misspecification: A well-known example is the assurance of normality by transforming the dependent variable (Sugasawa & Kubokawa, 2017; Tzavidis et al., 2018; Rojas-Perilla et al., 2019; Sugasawa & Kubokawa, 2019). Furthermore, the use of models under more flexible distributional assumptions is a fruitful approach (Diallo & Rao, 2018; Graf et al., 2019). From a different perspective, semi- or non-parametric approaches for the estimation of area-level means are investigated among others by Opsomer et al. (2008), using penalized spline components within the LMM setting. A distinct methodological option to avoid the parametric assumptions of LMMs are machine learning methods. These methods are not limited to parametric models and learn predictive relations from data, including higher order interactions between covariates, without explicit model assumptions (Hastie et al., 2009; Varian, 2014). Recently, Krennmair & Schmid (2022) introduce a framework enabling a coherent use of tree-based machine learning methods in SAE. They propose a non-linear, data-driven,

and semi-parametric alternative for the estimation of area-level means by using mixed effects random forests (MERF) in the methodological tradition of SAE. In general, random forests (RF) (Breiman, 2001) exhibit excellent predictive performance in the presence of outliers and implicitly solve problems of model-selection (Biau & Scornet, 2016). MERFs (Hajjem et al., 2014) combine these advantages with the ability to model hierarchical dependencies.

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

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

We demonstrate our methodology using the 2011 Socio-Economic Panel (SOEP) (Socio-Economic Panel, 2019) combined with aggregate census information from the same year to estimate the average individual opportunity cost of care work for 96 regional planning regions (RPRs) in Germany. We refer to care work as unpaid working hours attributed to child- or elderly-care reported by the SOEP. Opportunity cost is an economic concept comprising the time allocation problem, where the time allocated for care work implicitly corresponds to time not providing paid work (Buchanan, 1991). Informally provided care work has no direct corresponding monetary value and the determination of a correct shadow-price for the economic value is difficult. Classical interpretations of labour supply in economics such as Becker (1965) imply that an individual's hourly wage is an acceptable approximation to the unknown opportunity cost of time for working population. Thus,

we measure time cost by multiplying an individual's care time by the opportunity cost of the person's time represented as the reported hourly wage calculated also from reported income in the SOEP data. We are aware that our application is at best a first approximation making regional differences in opportunity cost of care work visible, accountable, and comparable. Unpaid care work mitigates public and private expenses on needed health services and infrastructure (Charles & Sevak, 2005). On the other hand, care-giving has a complex impact on the labour market (Truskinovsky & Maestas, 2018; Stanfors et al., 2019), for instance by affecting workforce individuals through personal or social burdens (Bauer & Sousa-Poza, 2015). From a macro-perspective, several studies examine the economic value of care work for countries through the concept of opportunity cost (Chari et al., 2015; Ochalek J., 2018; Mudrazija, 2019) and provide empirical evidence for policy measures.

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

The rest of the paper is structured as follows: Section 2.1 states a general mixed model that treats LMMs in SAE as special cases and enables the use of tree-based models. We consider the estimation of area-level means using MERFs, which effectively combine advantages of non-parametric random forests with the possibility to account for hierarchical dependencies. Section 2.2 describes our area-level mean estimator based on MERFs under limited data access. We scrutinize the use of EL calibration weights and subsequently address methodological limitations in Section 2.3. As a result, we propose a best practice strategy to ensure the proper usability of EL calibration weights in the context of SAE. Section 3 introduces a non-parametric bootstrap-scheme for the estimation of the area-level MSE. In Section 4, we use model-based simulations under complex settings to assess the performance of our stated methods for point and MSE estimates, showing that MERFs are a valid alternative to existing methods for the estimation of SAE means under limited data access. In Section 5, we estimate the average individual opportunity cost of care work for 96 RPRs in Germany using the 2011 SOEP data. After the introduction of data sources and direct estimates in Section 5.1, we highlight modelling and robustness properties of our proposed methods for point and uncertainty estimates compared to direct and other SAE estimates under limited auxiliary data. In Section 6, we conclude and

motivate further research.

2 Theory and Method

This section introduces a general mixed model enabling a simultaneous discussion of traditional LMM-based models in SAE such as the model of Battese et al. (1988) as well as semi-parametric interpretations such as the model of Krennmair & Schmid (2022) using MERFs. Section 2.2 provides details on our proposed methodology for MERFs under limited covariate data access and the determination of area-specific calibration weights based on EL. We close the section with a discussion on limitations of EL for SAE and state a best practice strategy ensuring the usability of our proposed point estimator in challenging empirical examples.

2.1 Model and Estimation of Coefficients

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

$$y_{ij} = f(\mathbf{x}_{ij}) + u_i + e_{ij} \quad \text{with} \quad u_i \sim N(0, \sigma_u^2) \quad \text{and} \quad e_{ij} \sim N(0, \sigma_e^2). \quad (1)$$

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

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

where $\bar{\mathbf{x}}_i = \frac{1}{N_i} \sum_{j \in U_i} \mathbf{x}_{ij}$ denotes area-specific population means on p covariates. In a variety of real-world examples, required assumptions for the BHF model hardly meet empirical evidence. Apart from transformation strategies to meet the required assumptions, non-parametric approaches can be used alternatively (Jiang & Rao, 2020). Tree-based machine learning methods such as RFs (Breiman, 2001) are data-driven procedures identifying predictive relations from data, including higher order interactions between

covariates, without explicit model assumptions (Hastie et al., 2009; Varian, 2014). RFs inherently perform model-selection and properly handle the presence of outliers (Biau & Scornet, 2016). However, an implicit assumption of tree-based models is the required independence of unit-level observations.

Defining f in Model (1) to be a RF results in a semi-parametric framework, combining advantages of RFs with the ability to model hierarchical structures of survey data using random effects. Krennmair & Schmid (2022) estimate area-level means with RFs (Breiman, 2001) introducing a method that enables the estimation of model-components $\hat{f}$, $\hat{u}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ in the context of SAE. The so-called mixed effects random forest (MERF) uses a procedure reminiscent of the EM-algorithm (Hajjem et al., 2014). For fitting Model (1) (where f is a RF) on survey data, the MERF algorithm subsequently estimates a) the forest function, assuming the random effects term to be correct and b) estimates the random effects part, assuming the Out-of-Bag-predictions (OOB-predictions) from the forest to be correct. OOB-predictions utilize the unused observations from the construction of each forest's sub-tree (Breiman, 2001; Biau & Scornet, 2016). The estimation of variance components $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_u^2$ is obtained implicitly by taking the expectation of ML estimators given the data. For further methodological details, we refer to Krennmair & Schmid (2022). The resulting estimator for the area-level mean for MERFs is summarized as:

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

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

2.2 MERFs under Aggregated Data

Estimates for the area-level mean μ_i using MERFs from Equation (2) require unit-level auxiliary census data as input for f. In contrast to the linear BHF model by Battese et al. (1988), aggregated covariate data cannot directly be used for non-linear or non-parametric procedures such as RFs, as in general $f(\bar{\mathbf{x}}_i) \neq \bar{f}_i(\mathbf{x}_{ij})$. Although the access to auxiliary population micro-data for the covariates imposes a limitation for practitioners, not many methods in SAE cope with the dual problem of providing robustness against model-failure, while simultaneously working under limited auxiliary data (Jiang & Rao, 2020). We propose a solution overcoming this issue by calibrating model-based estimates from MERFs in Equation (2) with weights that are based only on aggregated census-level covariates (means). The general idea originates from the bias-corrected transformed nested error regression estimator using aggregated covariate data (*TNER2*) by Li et al. (2019). We build on their idea of using calibration weights for SAE based on EL (Owen,

1990; Qin & Lawless, 1994; Owen, 2001) and transfer it to MERFs. As a result, our proposed method offers benefits of RFs such as robustness and implicit model-selection, while simultaneously working in cases of limited access to auxiliary covariate data. In short, our estimator for the area-level mean can be written as:

$$\hat{\mu}_i^{\text{MERFagg}} = \sum_{j=1}^{n_i} \hat{w}_{ij} \left[\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i \right]. \tag{3}$$

Note that optimal estimates for required model-components $\hat{f}$ and $\hat{u}_i$ are obtained similar to Equation (2) from survey data using the MERF algorithm as described by Krennmair & Schmid (2022). We incorporate aggregate census-level covariate information through the calibration weights w_{ij}, which balance unit-level predictions to achieve consistency with the area-wise covariate means from census data. Following Owen (1990) and Qin & Lawless (1994) the technical conditions for w_{ij} are to maximize the profile EL function $\prod_{j=1}^{n_i} w_{ij}$ under the following three constraints:

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

- $w_{ij} \geq 0$, ensuring the non-negativity of weights;

- $\sum_{j=1}^{n_i} w_{ij} = 1$, to normalize weights.

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

$$\hat{w}_{ij} = \frac{1}{n_i} \frac{1}{1 + \hat{\lambda}_i^\intercal (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i})}, \tag{4}$$

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

2.3 Limitation of Empirical Likelihood and a Best Practice Advice for SAE

The existence of an optimum solution to the maximization problem for the calibration weights $\hat{w}_{ij}$ is not necessarily guaranteed for applications in SAE. A necessary and sufficient condition ensuring the existence of a solution for $\hat{\lambda}_i$ is the existence of the zero vector as an interior point in the convex hull of constraint matrix $\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i}$. Especially for small sample sizes n_i this condition requires scrutiny (Emerson & Owen, 2009). If sample means of $\mathbf{x}_{ij}$ for area i strongly differ from $\bar{\mathbf{x}}_{\text{pop},i}$, for instance, due to a strong imbalance of individual sample values $\mathbf{x}_{ij}$ around the area-specific mean from population data $\bar{\mathbf{x}}_{\text{pop},i}$, no optimal solution for $\hat{\lambda}_i$ and subsequently $\hat{w}_{ij}$ can be obtained. The

dimensionality of existing covariates p relative to the sample size n_i exacerbates the problem. As a result, the constraints in matrix $\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i}$ are infeasible for finding a global optimum in Equation (4). Concrete empirical examples are different largely unbalanced categorical covariates in $\mathbf{x}_{ij}$, leading to column-wise multicollinearity in the $n_i \times p$ matrix of constraints $\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i}$.

Overcoming mentioned technical requirements, Li et al. (2019) propose the use of the adjusted empirical likelihood (AEL) approach by Chen et al. (2008), which forces the existence of a solution to Equation (4). Essentially, the introduced adjustment is an additional pseudo-observation within each domain i, increasing area-specific sample sizes to n_{i+1}. This pseudo-observation is jointly calculated from respective area-specific survey and census means of covariates (Chen et al., 2008). Although the added adjustment-observation reduces risks of numerical instabilities, it simultaneously imposes difficulties from an applied perspective of SAE. Emerson & Owen (2009) scrutinize the application of AEL in the context of multivariate population means, maintaining that the added pseudo-observation distorts the true likelihood configuration even for moderate dimensions of p in cases of low area-specific sample sizes n_i. Chen et al. (2008, p. 430) note, that the problem is mitigated if the semi-parametric model is correctly specified and if the initial estimates for $\bar{\mathbf{x}}_{\text{smp},i}$ are not too far away from the true population mean. Nevertheless, we observe that the influence of the bound-correction of Chen et al. (2008) used by Li et al. (2019) has drawbacks, which we will discuss in the model-based simulation in Section 4. Dealing with empirical examples characterized by low domain-specific sample sizes, we abstain from the approaches of adding synthetic pseudo-observations to each domain. We maintain that in the context of non-linear semi-parametric approaches (such as RFs) there is a risk of including implausible individual predictions from f based on the pseudo-covariates, i.e. $\hat{y}_{\text{pseudo},i}$. In this sense, pseudo-observations manipulate the estimation of area-level means under limited auxiliary information in two ways: indirectly through their effect on the determination of all weights $\hat{w}_{ij}$ and directly through the predicted pseudo-value that is added to the survey sample.

We postulate a stepwise approach to ensure a solution to Equation (4) for each area i under a reduced risk of distortions driven by improper pseudo-values through optimization bound-corrections. This approach can be interpreted as a best-practice strategy on the incorporation of maximal auxiliary covariate information through calibration weights in Equation (4) for the estimation of area-level means with MERFs. In detail, we first check for each area i whether perfect column-wise-dependence in the $p \times n_i$ matrix of constraints $(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i})_{j=1,\dots,n_i}$ exists. If so, we remove perfectly collinear columns and rerun the optimization. Subsequently, we proceed along two dimensions: a) increasing the sample size of i-th area and b) decreasing the number of auxiliary covariates p to calculate $\hat{w}_{ij}$ for area i. For a) we advise to sample a moderate number of observations (e.g., 10) randomly with replacement from an area which is "closest" to area i. We refer to

areas as "closest", if they have the smallest Euclidean distance in census-level information $\bar{\mathbf{x}}_{\text{pop},i}$. This additionally allows to handle out-of-sample areas. For b) we propose a backward selection of covariate information based on the variable importance. Variable importance are RF-specific metrics that enable the ranking of covariates reflecting their influence on the predictive model. As we are primarily concerned about the order of influence of covariates, we rank based on the mean decrease in impurity importance, which measures the total decrease in node-specific variance of the response variable from splitting, averaged over all trees (Biau & Scornet, 2016). Overall, our strategy to handle potential failure in the solutions for weights and out-of-sample domains is summarized in the following algorithmic strategy:

---

1. Use MERF to obtain estimates $\hat{f}$, $\hat{u}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ from available unit-level survey data and estimate the indicator $\hat{\mu}_i^{\text{MERFagg}}$ (3) including weights $\hat{w}_{ij}$ following Equation (4).

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

   (a) Check the feasibility of constraints used in the optimization and remove perfectly co-linear columns in $(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i})_{j=1,\dots,n_i}$. Retry the optimization in Equation (4).

   (b) If the calculation of weights fails again, optionally enhance the domain-specific sample size of area i by sampling randomly with replacement from the most "similar" domain according to the minimal row-wise Euclidean distance between area-specific aggregated covariate vectors $\bar{\mathbf{x}}_{\text{pop},i}$. Retry the calculation of weights $\hat{w}_{ij}$.

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

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

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

The general performance is illustrated by the results of the model-based simulation in Section 4. Furthermore, the proposed best-practice strategy will be demonstrated in the application in Section 5.

3   Uncertainty Estimation

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

General statistical results concerning the inference of area-level indicators from MERFs in SAE are rare, especially in comparison to the existing theory of inference using LMMs. Although the theoretical background for predictions from RFs grows (Sexton & Laake, 2009; Wager et al., 2014; Wager & Athey, 2018; Athey et al., 2019; Zhang et al., 2019), existing research mainly aims to quantify the uncertainty of individual predictions. From a survey perspective, Dagdoug et al. (2021) recently analyse theoretical properties of RF in the context of complex survey data. The extension of these results for partly-analytical uncertainty measures in the context of dependent data structures and towards area-level indicators is non trivial and a conducive topic for theoretical SAE.

In this paper, we propose a non-parametric bootstrap for finite populations estimating the MSE of the introduced area-level estimator under limited aggregate information defined by Equation (3). Essentially, we aim to find a solution to two problems simultaneously: Firstly, we need to flexibly capture the dependence-structure of the data and uncertainty introduced by the estimation of Model (1). Secondly, we face problems in simulating a full bootstrap population in the presence of aggregated census-level data.

Our proposed solution to this dual problem is the effective combination of two existing bootstrap schemes introduced by Chambers & Chandra (2013) and González-Manteiga et al. (2008). Addressing the problem of non-parametric generation of random components, we rely on the approach introduced by Chambers & Chandra (2013). One key-advantage is its leniency to potential specification errors of the covariance structure, as the extraction of the empirical residuals only depends on the correct specification of the mean behaviour function f of the model. Solving the problem of missing unit-level population covariate data, we base the general procedure on the methodological principles of the parametric bootstrap for finite populations introduced by González-Manteiga et al. (2008) adapted to the estimation of domain-level means. This allows us to find (pseudo-)true values by generating only error components instead of simulating full bootstrap populations. An important step concerning the handling and resampling of empirical error components is centring and scaling them by a bias-adjusted residual variance proposed by Mendez & Lohr (2011). In short, the estimator of the residual variance under the MERF from Equation

(2), $\hat{\sigma}_\epsilon^2$ <mark>is positively biased,</mark> as it includes excess uncertainty concerning the estimation of function $\hat{f}$. Further methodological details on the modification of the approach by Chambers & Chandra (2013) for MERFs for area-level means under unit-level models are found in Krennmair & Schmid (2022). Note that our proposed non-parametric MSE-bootstrap algorithm works for in- and out-of sample areas. The steps of the proposed bootstrap are as follows:

---

1. Use estimates $\hat{f}$, $\hat{\sigma}_e$, $\hat{\sigma}_u$, and respective weights $\hat{w}_{ij}$ from the application of the proposed method as summarized in Equation (3) on survey data with metric target variable y_{ij}.

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

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

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

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

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

   (b) Compute (pseudo-)true values for the population based on the fixed effects from area-wise mean estimates $\hat{\mu}_i^{\text{MERFagg}}$, as:

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

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

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

   We use OOB-predictions from $\hat{f}$ to imitate variations of $\mathbf{x}_{ij}$ covariates through

predictions from unused observations within each tree in the fitting process that vary throughout the bootstrap replications.

(d) Estimate $\hat{\mu}_i^{\mathrm{MERFagg}(b)}$ with the proposed method from Equation (3) on bootstrap sample values $y_{ij}^{(b)}$. Note that weights $\hat{w}_{ij}$ remain constant over B replications because the original survey covariates $\mathbf{x}_{ij}$ and population-level covariates $\bar{\mathbf{x}}_{\mathrm{pop},i}$ remain unchanged over B.

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

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

4   Model-Based Simulation

The model-based simulation allows for a controlled empirical assessment of our proposed methods for point and uncertainty estimates. Overall, we aim to show, that the proposed methodology from Section 2 and Section 3 performs as well as traditional SAE methods and has advantages in terms of robustness against model-failure. In particular, we study the performance of the proposed MERFs under limited data access (*MERFagg*, (3)) to the *direct* estimator, the *TNER2* estimator proposed by Li et al. (2019), the *BHF* estimator (Battese et al., 1988) as well as the MERF assuming access to unit-level census data (*MERFind*, (2)) by Krennmair & Schmid (2022). The *direct* estimator only uses sampled data to estimate the mean, which implies a strong dependence between the area-specific sample size and the quality of estimates. The *BHF* model serves as an established baseline model for the estimation of area-level means under limited auxiliary data. The *TNER2* aims to provide an alternative to the *BHF*, introducing aspects of transformations under limited data access. General differences in the performance of the *direct*, *BHF*, and *TNER2* estimator to the two MERF candidates (*MERFagg*, *MERFind*) indicate advantages of semi-parametric and non-linear modelling in the given data scenarios. The additional inclusion of the *MERFind* enables a direct comparison regarding the effect of access to aggregated auxiliary data (*MERFagg*) and existing unit-level auxiliary data (*MERFind*).

We consider four scenarios denoted as *Normal*, *Pareto*, *Interaction*, and *Logscale* and repeat each scenario independently $M = 500$ times. All four scenarios assume a finite population U of size $N = 50000$ with $D = 50$ disjunct areas $U_1, ..., U_D$ of equal size $N_i = 1000$. We generate samples under stratified random sampling, utilizing the 50 small areas as stratas, resulting in a sample size of $n = \sum_{i=1}^{D} n_i = 1229$. The area-specific

Table 1: Model-based simulation scenarios

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

sample sizes range from 5 to 50 sampled units with a median of 21 and a mean of 25. The sample sizes are comparable to area-level sample sizes in the application in Section 5 and can thus be considered to be realistic.

The choice of the simulation scenarios is motivated by our aim to evaluate the performance of the competing methods for economic and social inequality data. This includes skewed data, deviations from normality of error terms, or the presence of unknown non-linear interactions between covariates, that might trigger model-misspecifications in traditional SAE approaches based on LMMs. The data generating processes for the used scenarios are provided in Table 1. Scenario *Normal* provides a baseline under a LMM with normally distributed random effects and unit-level errors. As the model assumptions for LMMs are fully met, we aim to show that the *MERFagg* performs similarly well compared to linear competitors. Scenario *Pareto* is based on the same linear additive structure as scenario *Normal*, but has Pareto distributed unit-level errors. This leads to a skewed target variable, comparable to empirical cases of monetary data. The data generating process of scenario *Interaction* likewise results in a skewed target variable y_{ij}, although it shares its structure of random components with *Normal*. The *Interaction* scenario portrays advantages of semi-parametric and non-linear modelling methods protecting against model-failure arising from models with unknown interactions. Scenario *Logscale* introduces an additional example resulting in a skewed target variable. Log-normal distributed variables mimic realistic income scenarios and constitute a showcase for SAE transformation approaches. We want to show the ability of MERFs and particularly of *MERFagg* to handle such scenarios as well by identifying the non-linear relation introduced trough the transformation on the linear additive terms.

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

(RRMSE-RMSE):

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

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

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

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

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

where $\hat{\mu}_i^{(m)}$ is the estimated mean in area i based on any of the methods mentioned above and $\mu_i^{(m)}$ defines the true mean for area i in replication m. $MSE_{\mathrm{est},i}^{(m)}$ is estimated by the proposed bootstrap from Section 3.

For the computational realization of the model-based simulation, we use R (R Core Team, 2020). The *BHF* estimates are realized from the sae-package (Molina & Marhuenda, 2015). For the estimates of the *TNER2*, we used code provided by Li et al. (2019). For estimates based on the MERF approach, we use the packages ranger (Wright & Ziegler, 2017) and lme4 (Bates et al., 2015) to implement our method (*MERFagg*) and the *MERFind* estimator (Krennmair & Schmid, 2022). For RFs, we set the number of split-candidates to 1, keeping the default of 500 trees for each forest.

4.1 Performance of Point Estimators of the Small Area Means

We start with a focus on the performance of point estimates. Figure 1 reports the empirical RMSE of each point estimation method under the four scenarios. As expected, the *direct* estimates perform poorest due to the low sample sizes and the complexity of the data generating process. In these specific settings, the *TNER2* estimator outperforms *direct* estimates but performs worse compared to the *BHF*. In the *Pareto* and *Logscale* scenario, benefits of transformations might be suppressed by the influence of pseudo-observations due to the AEL approach, as discussed throughout the methodological Section 2.3 of this paper.

In the *Normal* scenario, the *BHF* performs best as it replicates the data generating process. The *MERFind* and the *MERFagg* perform on a comparable level, underlining the quality of our proposed calibration approach to incorporate aggregated census-level
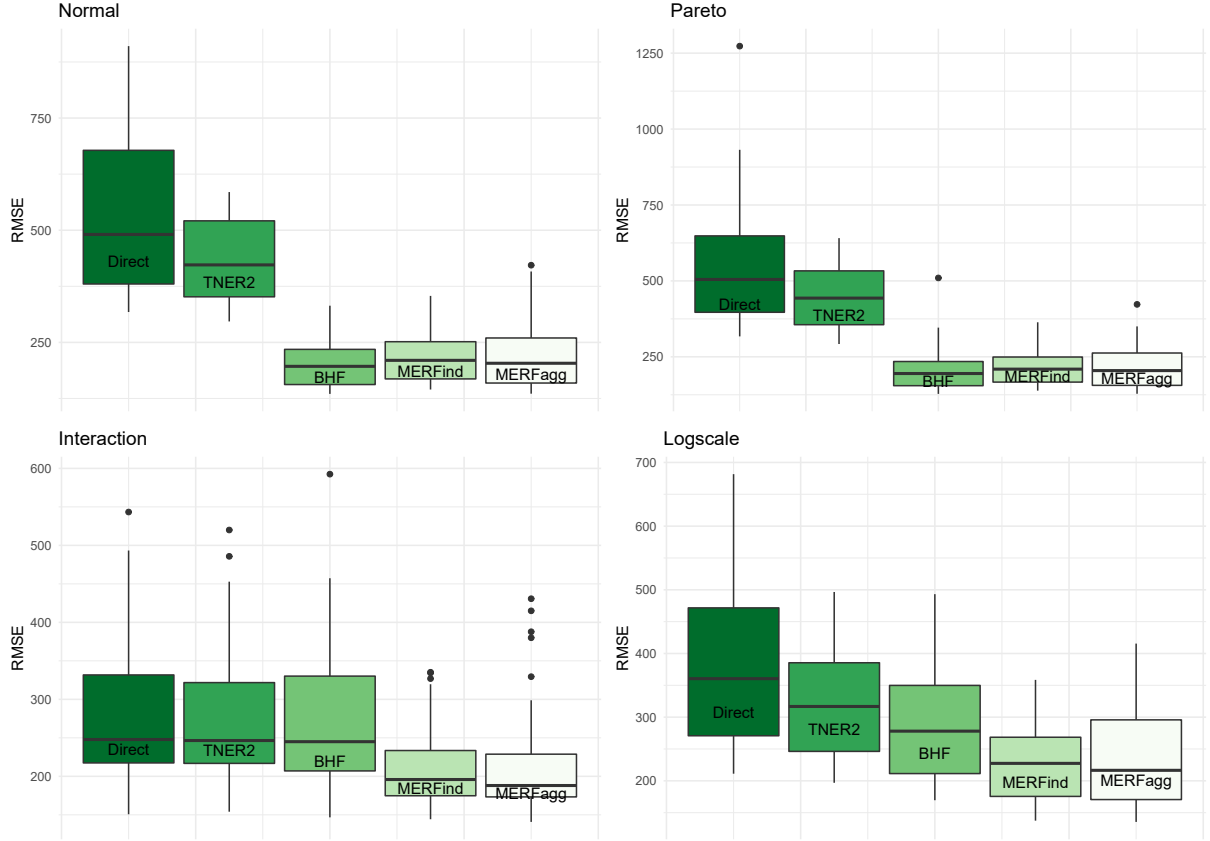
Figure 1: Empirical RMSE comparison of point estimates for area-level averages under four scenarios

information through the weights. *MERFagg* shows a better performance in median values, however the range of area-specific RMSE values is larger compared to MERF estimates based on unit-level census information. One area with particularly low sample size has a relatively high level of RMSE, which is explainable by the dependence of the optimum function for the weights in Equation (4) on n_i.

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

In the *Interaction* scenario, the point estimates of the proposed *MERFagg* outperform traditional SAE approaches under limited auxiliary information. Apparently the LMM-based methods cannot sufficiently capture the underlying predictive relation between the covariates, while the MERFs detect the non-linear term. Regarding the impact of restricted covariate data access, we observe relatively low values of mean and median RMSE compared to the hypothetical case of existing unit-level data in *MERFind*. Four outliers in areas with low sample sizes for *MERFagg* become apparent, although the median RMSE is lowest. We maintain, that this phenomenon can be mitigated if we increase the size

Table 2: Mean and Median of RB and RRMSE over areas for point estimates in four scenarios

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

of "close" observations from other areas to a higher level, especially in cases of complex interactions of effects in covariates such as *Interaction*.

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

Overall, the results from Figure 1 indicate that the MERF performs comparably well to LMMs in simple scenarios, and outperforms traditional SAE models in the presence of complex data generating processes, such as unknown non-linear relations between covariates or non-linear functions. Additionally, the robustness against model-misspecification of MERFs and their calibration weights $\hat{w}_{ij}$ holds if distributional assumptions for LMMs are not met, i.e. in the presence of non-normally distributed errors and skewed data. The influence of unit-level versus aggregated covariate information appears to be marginal in all of our four scenarios. We observe a moderate dependence between sample sizes and the quality of area-specific means for *MERFagg*, which is mainly explained by the way the calibration weights rely on the quality of survey data for a respective area i as discussed in Section 2.2.

Table 2 reports the corresponding values of RB and RRMSE for the discussed point estimates. The RB and the RRMSE from the *MERFagg* attest a competitively low level under all scenarios. All model-based MERF estimators have a lower mean and median RRMSE compared to the *direct* estimator in all scenarios. Despite a few outliers for RMSE and RB (cf. Figure 1), the median and mean values of *MERFagg* are remarkably low emphasizing the quality of estimates given the the substantial reduction in required covariate information.

4.2   Performance of the Bootstrap MSE Estimator

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

For each scenario and each simulation round, we choose $B = 200$ bootstrap replications. From the comparison of RB-RMSE among the four scenarios provided in Table 3, we infer, that the proposed non-parametric bootstrap-procedure effectively handles all four scenarios. This is demonstrated by relatively low mean values of positive RB-RMSE over the 50 areas after M replications. From an applied perspective, we prefer over- to underestimation for the MSE as it serves as an upper bound. We mainly use the area-level MSE for the further assessment in terms of CVs and consequently overestimation of area-level MSEs leads to an increased CVs. If our CVs are still below the thresholds, the estimates are definitely acceptable. The difference in RB-RMSE between *Normal* and *Pareto* is marginal, indicating that the non-parametric bootstrap effectively handles non-Gaussian error terms.

Table 3: Performance of MSE estimator in model-based simulation:
mean and median of RB-RMSE and RRMSE-RMSE over areas

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

Figure 2 provides additional intuition on the quality of our proposed non-parametric MSE-bootstrap estimator. Given the area-wise tracking properties in all four scenarios, we conclude that our MSE estimates strongly correspond to the empirical RMSE. We infer that the overestimation in Table 3 is mainly driven by overestimation in areas with low sample sizes. Thus, our non-parametric MSE estimator provides an upper bound for the uncertainty of particular difficult point estimates due to low sample sizes. Apart from this characteristic, we observe no further systematic differences between the estimated and empirical MSE estimates regarding their performance throughout our model-based simulation.
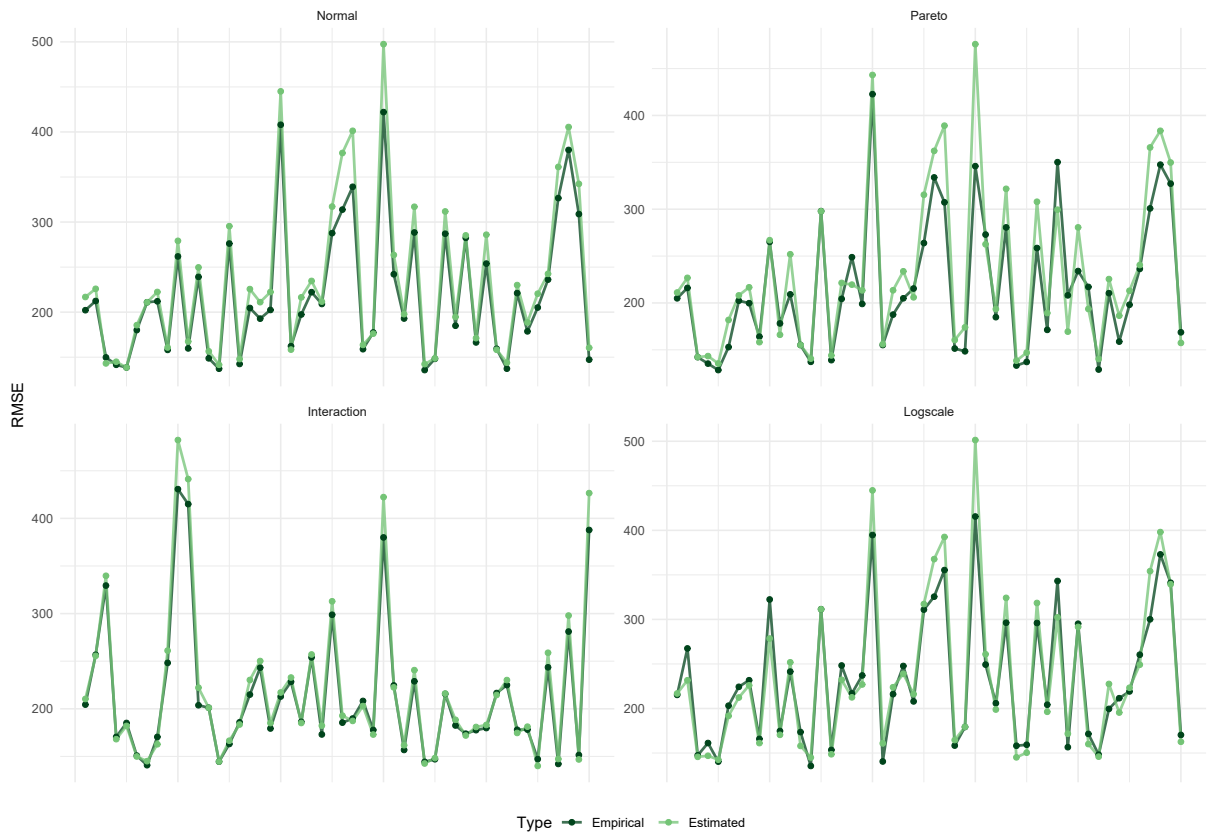
Figure 2: Estimated and empirical area-level RMSEs for four scenarios

5 Application

This section starts with a description of data sources and outlines our empirical analysis. We describe the survey data SOEP (Socio-Economic Panel) and discuss primary *direct* estimates on spatial differences of average individual opportunity cost of care work for German RPRs. Moreover, we propose the use of model-based SAE, which incorporates auxiliary variables from the 2011 German census. Demonstrating our proposed method of MERFs with aggregated data for point and uncertainty estimates, we show advantages to existing model-based SAE methods. Finally, we discuss our empirical findings concerning the cost of care work in Germany. We conduct the analysis with R (R Core Team, 2020).

5.1 Data Sources and Direct Estimates of Spatial Opportunity Cost of Care Work

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

We construct the target variable of individual monthly opportunity cost of care work from the SOEP in 2011 (Socio-Economic Panel, 2019) and use the available refreshment samples. We choose the year 2011 because the last census was in this year and therefore census and survey data have no time inconsistencies. The underlying sampling design is a multi-stage stratified sampling procedure: Initially, stratification is carried out into federal states, governmental regions, and municipalities. Subsequently, addresses are sampled using the random walk methodology within each primary sampling unit (Kroh et al., 2018). Our analysis focuses on the working age population aged between 15 to 64, as defined by international standards (OECD, 2020). In detail, we calculate the individual opportunity cost in Euro per month for 2011 as follows: first, we compute opportunity cost as hourly wage by taking the mean gross individual income divided by hours of paid work. Then, we multiply the hours of monthly unpaid work due to child- or elderly-care by the hourly cost of opportunity. The resulting metric target variable y_{ij} for Germany is highly skewed, ranging from 0€ to 2413.79€ (mean: 100.96€ and median: 176.93€). A histogram is provided in Figure 3.

In total we have 3939 sample survey observations. National averages do not serve for
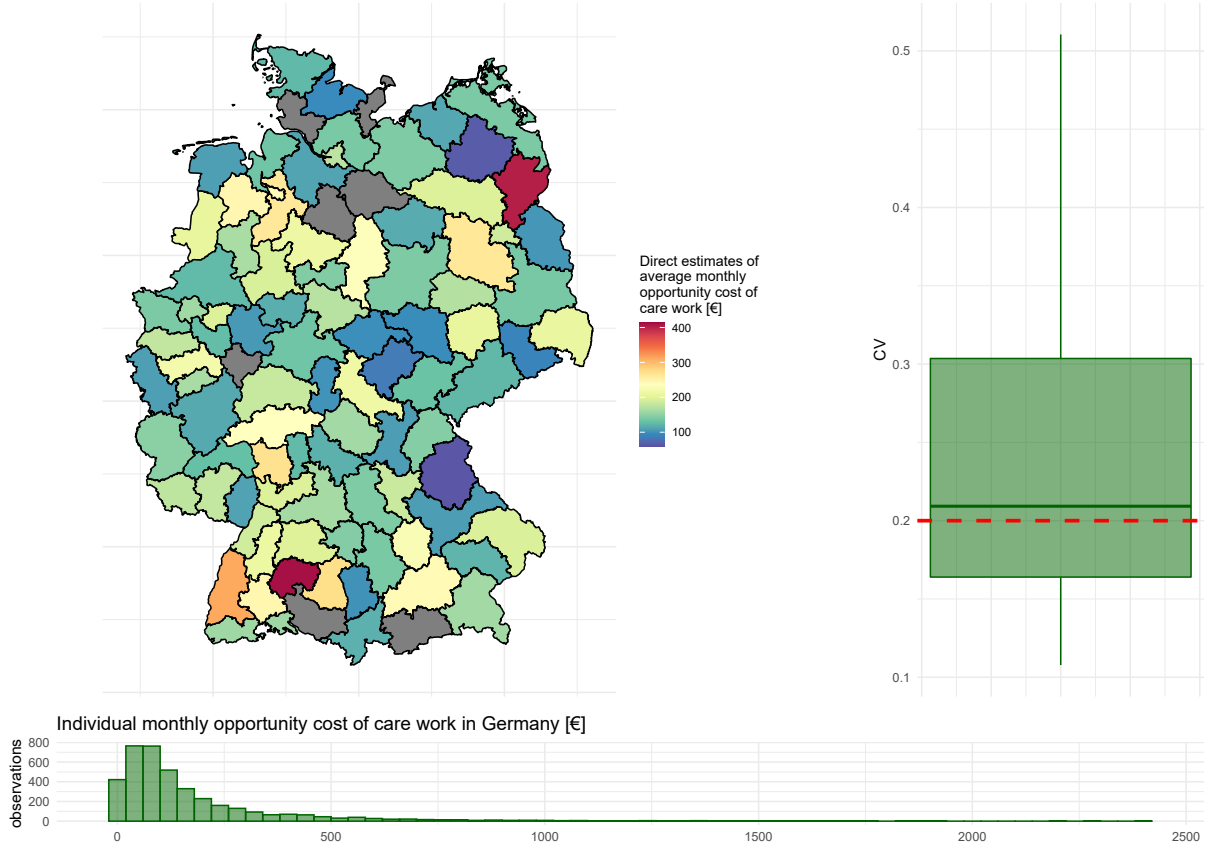
Figure 3: Overview of *direct* estimates, corresponding CVs and the distribution of opportunity cost of care work in Germany.

monitoring efficacy of regional developments and policy measures. Our major interest is a finer spatial resolution to map regional patterns of opportunity cost of care work across Germany. We analyse 96 respective RPRs in Germany, resulting in area-specific sample sizes from 4 to 158 with a mean of 35 and median of 41. First results of *direct* estimates can be seen in the map in Figure (3). Estimates of the mean monthly opportunity cost of individual care work range from 64.31€ (Oberpfalz-Nord) to 409.38€ (Neckar-Alb). In general, we observe no major difference between former East and West Germany. Additionally, levels of opportunity cost are higher in metropolitan areas surrounding cities than in the cities itself and compared to rural areas.

Small sample sizes lead to unreliable estimates accompanied by high variances. Furthermore, we are not allowed to report *direct* estimates from regions with sample size below 10 due to confidentiality agreements with the data provider. This is the case for 7 RPRs. To obtain variances and subsequently determining the coefficients of variation (CV) for the *direct* estimates, we use the calibrated bootstrap by Alfons & Templ (2013) implemented in the R-package emdi by Kreutzmann et al. (2019). Eurostat (2019) postulates that estimates with a CV of less than 20% can be considered as reliable. As reported by Figure 3, more than half of the regions (47 out-of reaming 89) exceed this threshold.

The *direct* estimation results suffer from differences in quality due to low area-level sample

sizes and specifically high variability. Model-based SAE methods help to improve the estimation accuracy of results. As SOEP auxiliary variables are measured in the same way as in the Germans census (Statistisches Bundesamt, 2015), census covariate data can serve as auxiliary information needed in SAE models. However, the German census provides information only at aggregated RPR-levels. Overall, we have 19 covariates on personal and socio-economic background within our sample for which we additionally received corresponding means from the German Statistical Office calculated from the German 2011 census. Details on available covariates and their variable importance is provided within the Appendix in Table 4.

5.2   Model-Based Estimates

This section illustrates the application of our proposed method for MERFs with aggregate covariate data for the estimation of area-level means. We map the estimated monthly mean opportunity cost of unpaid care work for 96 RPRs in Germany for the year 2011. Moreover, we assess the quality of our estimates by providing CVs based on our proposed non-parametric MSE-bootstrap procedure discussed in Section 3 and juxtapose our results to the previously discussed *direct* estimates and the well-established BHF model by Battese et al. (1988). A full comparison to the *TNER2* estimates (Li et al., 2019) is not possible because Li et al. (2019) do not provide uncertainty estimators required for a qualitative comparison in terms of CVs.

As reported by Figure 3, our target variable of individual opportunity cost is highly skewed, indicating that traditional LMMs (such as the BHF) run the risk of model-misspecification. In contrast, our proposed procedure shows robustness against model-failure due to outliers or complex data structures. Apart from specifying separate regions being modelled as random intercepts, the proposed *MERFagg* approach can be seen as purely data-driven: We train a predictive model on the survey set and incorporate as much auxiliary information for the determination of area-specific calibrations weights as possible based on the variable importance obtained from the fitted RF object $\hat{f}$. For this example we set the tuning parameter of the RF to 500 sub-trees. Repeated $5-$fold cross-validation supports the choice of proposing 5 randomly drawn split candidates at each split for the forest. Regarding our best-practice strategy, we chose that we want to calculate the weights based on a minimum of the 3 most influential variables. An overview of the number of covariates included can be found in the appendix (Figure 7). For the non-parametric MSE bootstrap-procedure, we use $B = 200$.

The results from the application of *MERFagg* are reported in Figure 6. We primarily focus on a discussion of technical details of estimates from our proposed approach and postpone the contextual discussion of results to the end of this section. Overall we observe a dominance of covariates of age, size of the household, households with a child, gender
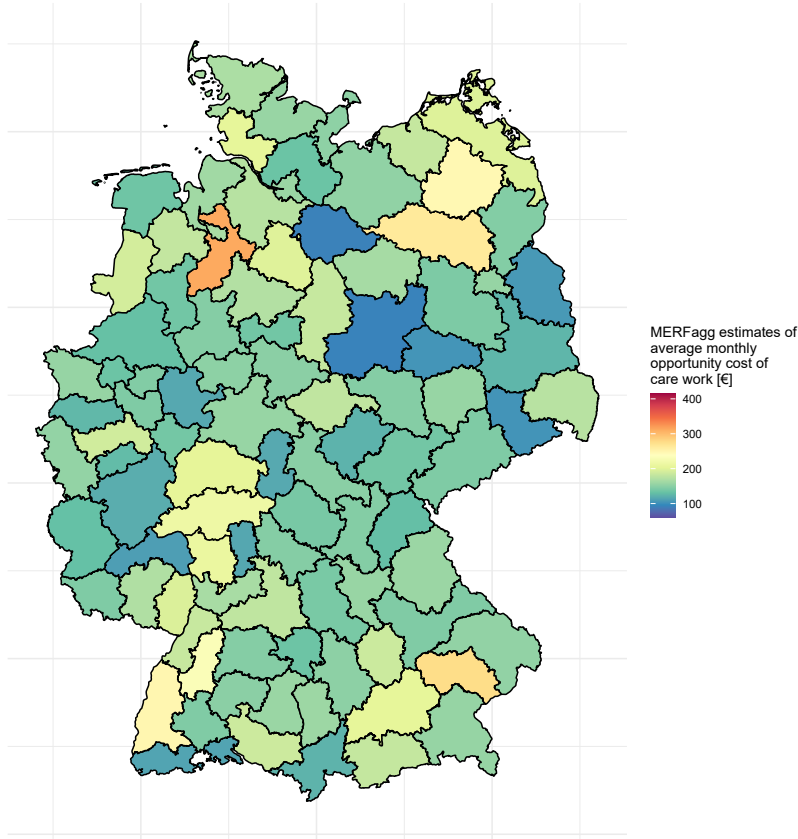
Figure 4: Spatial representation of area-level mean estimates from
MERFagg (3) for mean monthly opportunity cost of care work [€].

and whether the person is employed in the public sector (cf. Table 4 in the Appendix).
Throughout all 96 areas, we incorporate auxiliary information from 3 up to 15 covariates
from census-level aggregates through optimal calibration-weights $\hat{w}_{ij}$. A detailed map on
the number of included census-level covariates is provided in the Appendix within Figure
7. Unfortunately this attempt failed for 5 regions, which were left with uninformative
weights $\hat{w}_{ij} = 1/n_i$. Although these estimates do not incorporate auxiliary information,
recall from Equation (3) that the corresponding estimates are reduced to $\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i$ and
thus still rely on the model-based estimates comprising information from other in-sample
areas.

A comparison between the maps from *direct* estimates in Figure 3 and estimates based on
MERFagg from Figure 4 indicates that results form *MERFagg* appear to be more balanced
and overall no major differences regarding changes in regional patterns of opportunity cost
of care work are observable. Figure 5 sorts areas by increasing survey sample sizes and
thus allows for a more precise discussion on peculiarities of point estimates for area-level
means of monthly opportunity cost for the 96 RPRs. Estimates from the BHF method
are produced form the R-package sae (Molina & Marhuenda, 2015). Although, the raw
comparison of point estimates only allows for limited findings regarding the quality of
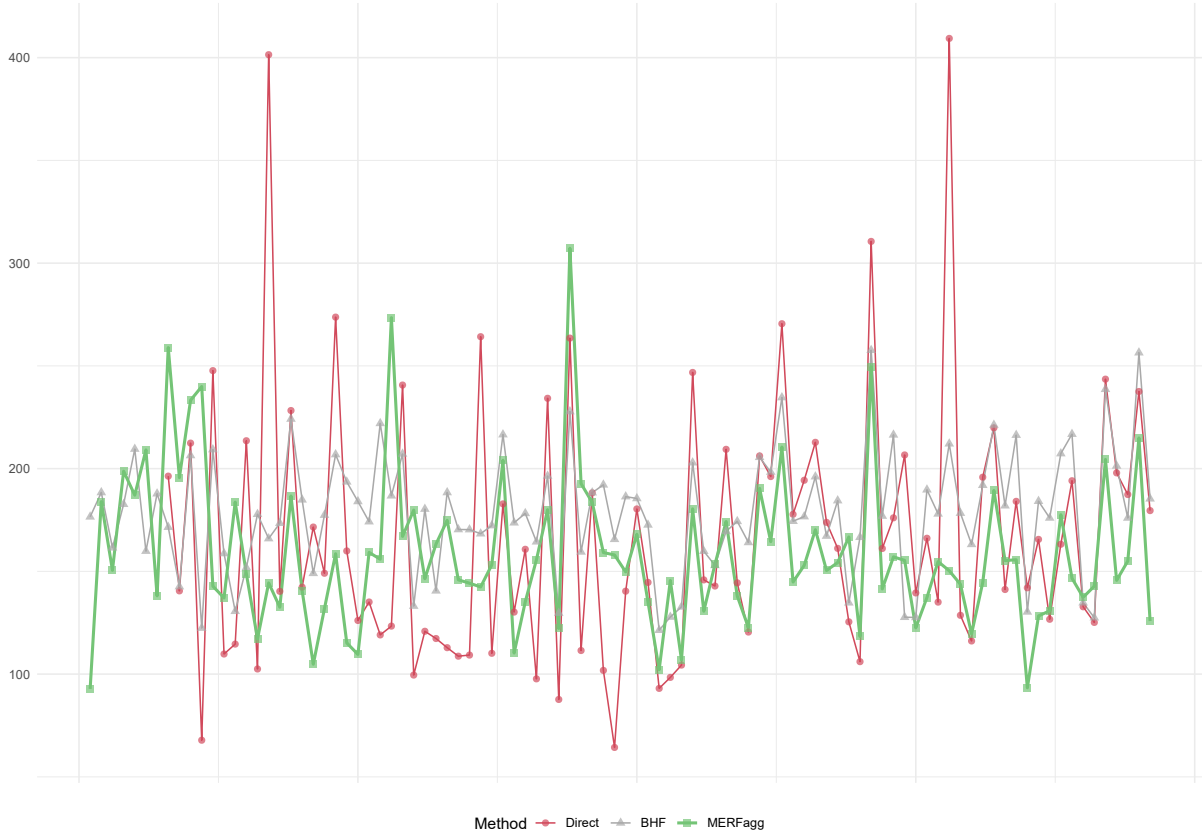methods, we report the mitigation of two outlier-driven direct estimates. Compared to

22

Figure 5: Detailed comparison of area-level mean estimates for monthly opportunity cost of care work [€]. The 96 German RPRs are sorted by increasing sample size. We compare results based on methods *direct*, *BHF*, and *MERFagg*.

the *direct* estimates, as well as the estimates from the BHF, the *MERFagg* produces relatively lower values although the estimates track patterns of high- and low levels with increasing survey sampling size.

As already discussed, *direct* estimates suffer from relatively low accuracy measured by their respective CVs. Figure 6 juxtaposes CVs for *direct* estimates, the *BHF*, and our proposed method of *MERFagg* to contextualize the performance of point estimates from Figure 5. We observe that CVs for *MERFagg* are on average smaller compared to CVs from *direct* estimates as well as the BHF. According to the boxplots in Figure 6, model-based estimates produce more accurate results indicated by lower CVs than *direct* estimates. *MERFagg* shows the lowest CVs compared to the other methods in mean and median-terms. Two areas can be considered as outliers reporting CVs over 0.3. For one of these two regions, the calculation of weights failed. The *MERFagg* estimates improve the *direct* estimates: Only 15 areas from 96 do not meet the required threshold of 20%. As expected, especially for areas that are unreliable due to low sample sizes, model-based estimates improve the accuracy. In turn, we observe that the *direct* estimates are relatively accurate for areas with high sample sizes. Compared to other model-based SAE methods, survey
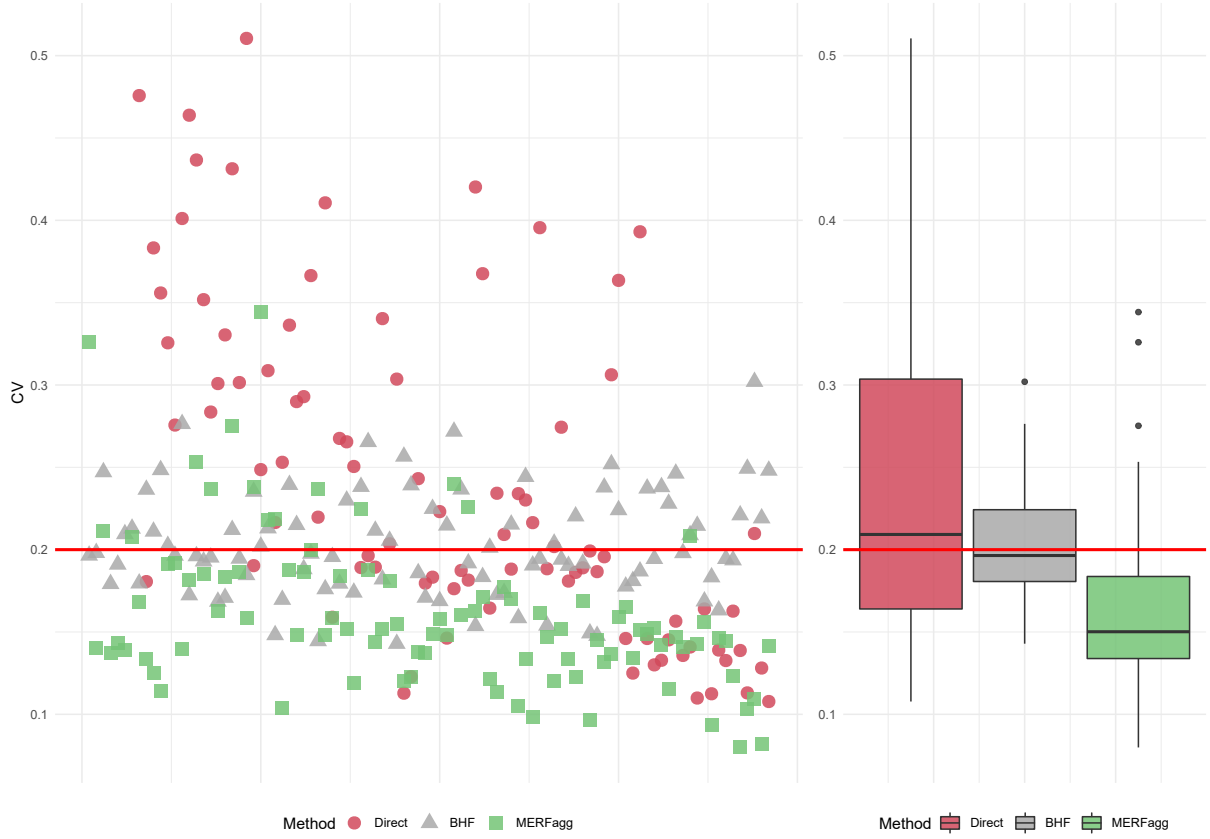
Figure 6: Left: Comparison of area-specific CVs ordered from low to high sample sizes. Right: Comparison of CVs over 96 respective areas between direct, *BHF* and *MERFagg*. The red line marks the 20%-criterion for defining reliable estimates by Eurostat (2019).

weights are not directly used in the model-fitting for *MERFagg*. Although it is generally possible to incorporate survey weights in the importance sampling within a forest, we maintain that the efficient use of survey weights with MERFs for the estimation of area-level indicators requires further research which would exceed the scope of this paper.

Overall, all RPRs throughout Germany report comparable levels of average individual monthly opportunity cost of care work. Nevertheless, a detailed inspection of Figure 4 reveals a small cluster of lower values in the North-East of Germany. From a causal perspective, the explanation of such patterns appears to be difficult and not effective. Wage and individual opportunity cost directly relate while time spent for care work negatively affects opportunity cost. Thus, it is not observable whether the effect is driven by differences in average income or increased time-allocation for care work or both. On the other hand, the concept allows us to uncover and map the value of unpaid care work on a sub-regional-level in Germany.

24

6 Conclusion

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

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

the unobservable "real" shadow price.

We motivate two major dimensions for further research, including theoretical work and aspects of generalizations. From a theoretical perspective, further research is needed to investigate the construction of a partial-analytical MSE for area-level means or the construction of an asymptotic MSE estimator. From a statistical perspective, an in-depth analysis regarding the effects of incorporating survey weights into RFs and particularly MERFs under aggregated covariate data is needed for point and uncertainty estimates, as this would clearly exceed the scope of the present paper. Our approach shares the EL-calibration-argument with Li et al. (2019), however, saves on the computationally intensive procedure of a smearing step (Duan, 1983) without drawbacks on the predictive performance, because no transformations and corresponding bias exists. Nevertheless, we maintain that pairing our approach with a smearing argument allows for a more general methodology and subsequently for the estimation of indicators such as quantiles (Chambers & Dunstan, 1986). Although, we will leave a detailed discussion of this idea to further research, a short outline of the argument can be found in the Appendix 7.2. Apart from generalizations to quantiles, the approach of this paper is generalizable to model (complex) spatial correlations. Additionally, a generalization towards binary or count data is possible and left to further research. The semi-parametric composite formulation of Model (1) allows for f to adapt any functional form regarding the estimation of the conditional mean of y_{ij} given x_{ij} and technically transfers to other machine learning methods, such as gradient-boosted trees or support vector machines.

7 Appendix

7.1 Additional Information on the Application (Section 5)

Table 4: Auxiliary variables on personal and socio-economic background and their variable importance based on the trained RF $\hat{f}$.

| Covariates	Variable importance
Age in years	30715147.623
Number of persons living in household	17109846.300
Position in Household: Child	7519805.884
Sex	4031803.086
Employment status: civil servants	3704520.439
Employment status: employed without national insurance (e.g. mini-jobber)	3078656.890
Tenant or owner	2632970.858
Position in Household: single parent	2500261.812
Migration background: direct	2453187.125
Position in Household: living alone	1380917.681
Position in Household: marriage-like	1341933.482
Migration background: indirect	1207604.491
Grouped nationality: European Union (excluding Germany)	697919.972
Grouped nationality: remaining European countries	468653.092
Grouped nationality: Asia	367207.174
Grouped nationality: North America	224042.331
Grouped nationality: Australia	45084.788
Grouped nationality: Africa	10109.844
Grouped nationality: South America	5150.957
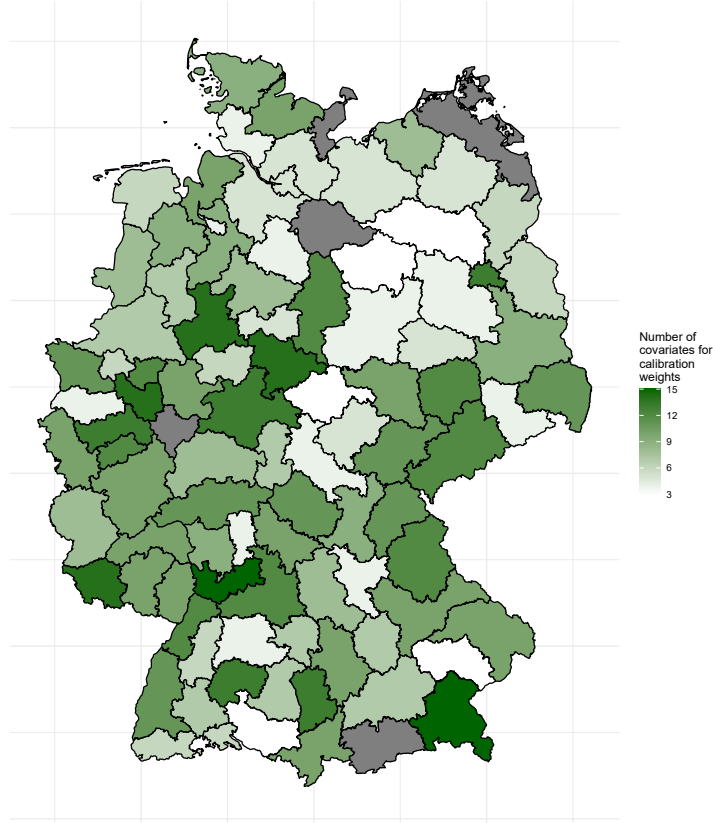
Figure 7: Inclusion of covariates through weights

7.2 Extension towards the Estimation of Quantiles

Smearing Approach and Estimation of Means: The smearing argument form Duan (1983) could be optionally inserted in Equation (3) to estimate mean values

$$\hat{\mu}_i^{\text{MERFagg Smearing}} = \sum_{j=1}^{n_i} \left[\hat{w}_{ij} \frac{1}{R} \sum_{r=1}^{R} (f(\mathbf{x}_{ij}) + \hat{u}_i + e_{ir}^*) \right], \tag{5}$$

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

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

Note that the formulation of Equation (5) coincidences with the estimator of Li et al. (2019), if we choose $f = \mathbf{x}_{ij}^{\mathsf{T}} \beta$ and draw e_r^* from $N(0, \hat{\sigma}_e^2)$. Additionally, they apply a data-driven transformation on $f(\mathbf{x}_{ij}) + \hat{u}_i + e_{ir}^*$.

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

$$\hat{F}_i^*(t) = N_i^{-1} \left[\sum_{j \in s_i} I(y_{ij} \leq t) + R^{-1} \sum_{j \in s_i} \sum_{r=1}^{R} n_i \hat{w}_{ij} I \left(\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i + e_{ir}^* \leq t \right) \right], \qquad (6)$$

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

The area-level quantile $q(i, \phi)$ of $\phi \in [0, 1]$ can straight forwardly be calculated by:

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

References

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

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

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

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

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

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

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

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

Buchanan, J. M. (1991). Opportunity cost. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *The world of economics* (pp. 520–525). London: Palgrave Macmillan UK.

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

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

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

Charles, K. K., & Sevak, P. (2005, November). Can family caregiving substitute for nursing home care? *Journal of health economics*, *24*(6), 1174—-1190.

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

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

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

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

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

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

Emerson, S., & Owen, A. (2009). Calibration of the empirical likelihood method for a vector mean. *Electron. J. Statist*, *3*, 1161—1192.

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

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

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

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The german socio-economic panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, *239*(2), 345–360.

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

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

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

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

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

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

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

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

Krennmair, P., & Schmid, T. (2022). *Flexible domain prediction using mixed effects random forests.* Available from `https://arxiv.org/pdf/2201.10933`.

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

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

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

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

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

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

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

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

OECD. (2020). *Working age population (indicator). OECD, Paris.* (Available from `https://www.oecd-ilibrary.org/social-issues-migration-health/working-age-population/indicator/english_d339918b-en`).

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

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

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

Owen, A. (2001). *Empirical likelihood.* Chapman and Hall/CRC.

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

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

Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2.nd ed.). New Jersey: Wiley: Wiley series in survey methodology.

R Core Team. (2020). R: a language and environment for statistical computing [Computer software manual]. Vienna.

Rojas-Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N. (2019). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(1), 121-148.

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

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

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

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

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

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

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

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

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., & Rojas-Perilla, N. (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.

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

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

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

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

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