The Survey Statistician, **2022, Vol. 86, 22–31.**

New and Emerging Methods

Tree-Based Machine Learning in Small Area Estimation

Patrick Krennmair [1], Nora Würz [2] and Timo Schmid[3]

[1]Freie Universität Berlin, Germany, patrick.krennmair@fu-berlin.de
[2]Freie Universität Berlin, Germany, nora.wuerz@fu-berlin.de
[3]Otto-Friedrich-Universität Bamberg, Germany, timo.schmid@uni-bamberg.de

Abstract

Reliable estimators of the spatial distribution of socio-economic indicators are essential for evidence-based policy-making. As the accuracy of direct estimates from survey data decrease with spatially finer target levels, small area estimation approaches are promising. In this article, we outline new approaches that combine small area methodology with machine learning methods. The presented semi-parametric approach is promising as it avoids the assumptions of linear mixed models in contrast to classical small area models and builds on random forests. These tree-based machine learning predictors have the advantage of robustness against outliers and implicit model-selection. As for classical small area models, we account for hierarchically dependent data. We present point estimators applicable to full as well as aggregated auxiliary data access and outline their uncertainty measure. We compare methods based on a reproducible and illustrative example using open-source income data from Austria.

Keywords: Official statistics; Mean squared error; Tree-based methods; Prediction

1 Introduction

Evidenced-based policy decisions require a solid and transparent empirical basis. An effective way to produce empirical findings is the construction of the target indicator using sampled information from individual and household surveys. Typically, we can partition a population into geographic, social, or political sub-units that are referred to as 'domains' or 'areas', which allows for the additional perspective of the spatial distribution of targeted indicators. Due to cost and efficiency constraints, the survey sample size is limited and at high spatial resolution the sample size within a domain might become small or even zero. Direct indicator estimates only use existing domain-level survey information. The implicit reduction of area-specific sample sizes as the level of required detail increases, leads to unreliable and imprecise direct estimates. A methodology that provides reliable and detailed estimates for this particular challenge is referred to as Small Area Estimation (SAE) (Pfeffermann, 2013; Rao & Molina, 2015; Tzavidis et al., 2018).

Model-based SAE methods improve estimates by linking survey data and available secondary auxiliary information (e.g. census or administrative data) via predictive models. This combination of information increases the effective sample sizes and subsequently the precision of domain-specific estimates. We broadly divide, SAE models in two classes: Area-level models - e.g Fay-Herriot models (Fay & Herriot, 1979) assuming aggregated data for survey and auxiliary information - and unit-level models - e.g. the nested error regression model by Battese et al. (1988) requiring access to a micro-level survey (Pfeffermann, 2013). Unit and area-level models alike are regression based and the hierarchical structure of observations is modelled by random effects. As a result, most of the SAE models are rooted within the methodological paradigm of Linear Mixed Models (LMM). Under the parametric framework, optimality estimators (under the assumed model) can be obtained. Jiang & Rao (2020) remind that model-based estimates follow the implied distribution of the model and predictive performance and inferences become erroneous and biased in cases of severe violations of model assumptions.

Working with social and economic datasets, we face heavily skewed and unbalanced target variables and models have to identify complex and indistinct relations between covariates. One strategy to prevent model-failure, is the assurance of normality by transforming the dependent variable improving the performance of unit-level models using a fixed logarithmic (Berg & Chandra, 2014; Molina & Martín, 2018) or data-driven (Sugasawa & Kubokawa, 2019; Rojas-Perilla et al., 2020) transformations. In cases of limited access to auxilliary information (i.e area-level aggregates of covariates from population data), small area means can be determined using robust methods like robustified linear mixed models (Sinha & Rao, 2009) or M-quantile based methods (Chambers & Tzavidis, 2006; Marchetti & Tzavidis, 2021). Another alternative is the use of models with less restrictive (parametric) assumptions to avoid model-failure. For instance, Diallo & Rao (2018) and Graf et al. (2019) formulate unit-level models under more flexible distributional assumptions. Semi- or non-parametric approaches for the estimations of area-level means were investigated among others by Opsomer et al. (2008). They use penalized splines regression, treating the coefficients of spline components as additional random effects within the LMM setting.

Machine Learning methods represent a further methodological option to avoid parametric assumptions of LMMs. 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; Gelman & Vehtari, 2021). Despite existing conceptual differences between machine learning and 'traditional' statistical methods (e.g. best possible predictions vs. parametric representation and interpretation), machine learning methods became a substantial element in statistical methodology research (Efron, 2020). For instance, the training/test-set paradigm is central to machine learning and conceptually transfers to the methodology of unit-level SAE-models: the survey data serves as a training-set to construct a proper model, while supplementary data (usually census, register or administrative data) of auxiliary information is used to predict final indicators over sampled and non-sampled areas. Jiang & Rao (2020) observe that SAE research is susceptible to novel approaches from various fields of statistics, however, results from machine learning are still harder to be interpreted and justified by SAE-practitioners compared to LMM-alternatives. Especially for SAE, new methods must meet the premise of basic principles of survey and inference theory. In this sense, the objectives of SAE coincide with the general perspective of Efron (2020), maintaining that an opportunity for modern statistics lies in the critical analysis and assessment of properties of predictive algorithms to make them 'scientifically applicable'. With this paper and our research, we aim to contribute to this purpose for SAE.

Among the broad class of machine learning methods, we focus on random forests (RFs) (Breiman, 2001) because they exhibit excellent predictive performance in the presence of complex and non-linear interactions and implicitly solve problems of model-selection (Biau & Scornet, 2016). The general idea of applying tree-based methods in SAE is not entirely new (Anderson et al., 2014; Bilton et al., 2017; De Moliner & Goga, 2018; Mendez, 2008). Recently, Dagdoug et al. (2021) analyse theoretical properties of RF in the context of complex survey data for model-assisted estimation. Krennmair & Schmid (2022) provide a consistent framework enabling a coherent use of tree-based machine learning methods in SAE and propose a non-linear, data-driven, and semi-parametric alternative for the estimation of area-level means using RFs in the methodological tradition of SAE. We will refer to this methodology combining the mixed effect model with RFs in the following as Mixed Effects Random Forest (MERFs). Section 2 introduces a general mixed effects model for SAE and its combination with RFs. Accordingly, the estimation of corresponding model-coefficients is explained and the MERF methodology to obtain domain-specific mean-estimates under unit-level and aggregated census information is elaborated in more depth. In addition, we outline the possibility of estimating the uncertainty of domain-specific indicators measured by corresponding mean squared errors (MSEs) in Section 2.3. An illustrative example on Austrian income data in Section 3 demonstrates both estimators from the theory part. Section 4 concludes and provides an outlook on further perspectives of research regarding the diversification of the model-toolbox for SAE-practitioners and researchers.

2 Using mixed effects random forests in SAE

RFs captivate with a lack of assumptions such as linearity or the distributional specification of model errors. Major benefits are the detection of higher order interactions between covariates, implicit model-selection, and the proper handling of outliers and high-dimensional covariate data without model assumptions (Hastie et al., 2009; Biau & Scornet, 2016). However, observations are assumed to be independent. Applications of SAE are characterized by the use of hierarchical data. Ignoring the correlation between observations, generally results in inferior point-predictions and inferences. Krennmair & Schmid (2022) introduce a general mixed model framework enabling the estimation of data-driven RFs, while simultaneously accounting for structural dependencies of survey data. This general formulation treats traditional LMM-based models in SAE as special cases and thus allows for a simultaneous discussion of existing SAE methods.

2.1 A general mixed effects model for SAE and MERFs

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 $\mathbf{x} = (x_1, x_2, ..., x_p)^\intercal$ includes p explanatory variables and is available for every unit j within the sample s. The relationship between $\mathbf{x}_{ij}$ and y_{ij} is assumed to follow a general mixed effects regression model:

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

Function $f(\mathbf{x}_{ij})$ models the conditional mean of y_{ij} given $\mathbf{x}_{ij}$. Area-specific random intercepts u_i account for the hierarchical dependency structure of observations and we subsequently assume unit-level errors e_{ij} and random effects u_i to be independent.

For instance, defining $f(\mathbf{x}_{ij}) = \mathbf{x}_{ij}^\intercal \beta$, where $\beta = (\beta_1, ..., \beta_p)^\intercal$, coincides with the well known nested error regression model proposed by Battese et al. (1988). This widely used LMM with area-specific random effects forms the basis for further unit-level SAE-models, such as the EBP (Molina & Rao, 2010) or the EBP under data-driven transformation by Rojas-Perilla et al. (2020). If the assumptions of the LMMs are met, optimal estimates of fixed effects $\hat{\beta}$ and variance components $\hat{\sigma}_u^2, \hat{\sigma}_e^2$ are obtained by maximum likelihood (ML) or restricted maximum likelihood (REML) (Rao & Molina, 2015).

If we assume f in Model (1) to be a RF (Breiman, 2001), we result in a semi-parametric framework, combining predictive advantages of RFs with the ability to model hierarchical structures of survey data using random effects. The method obtains optimal estimates of model components $\hat{f}$, $\hat{u}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ based on a procedure which is reminiscent of the EM-algorithm (Hajjem et al., 2014). The proposed MERF algorithm fits optimal parameters for Model (1) (where f is a RF) by iteratively estimating a) the forest function, assuming the random effects term to be correct and b) the random effects part, assuming the Out-of-Bag-predictions (OOB-predictions) from the forest to be correct. OOB-predictions utilize the unused observations from the construction of each forest's sub-tree (Breiman, 2001; Biau & Scornet, 2016). The estimation of variance components σ_ϵ^2 and σ_u^2 is obtained implicitly by taking the expectation of ML estimators given the data. The marginal change of a generalized log-likelihood criterion of the composite model monitors the convergence of the estimation algorithm. For further methodological details, we refer to Krennmair & Schmid (2022). The resulting estimator for model-based predictions from the MERF is summarized as follows:

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

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

Under the assumed existence of unit-level (i.e. $\mathbf{x}_{ij}$) population data (usually census or administrative data), $\hat{\mu}_{ij}^{\text{MERF}}$ in Equation (2) can predict conditional means of a metric dependent variable. As typical for SAE, our major interest is in estimating area-level means. The domain-level mean estimator for each area i is given by:

$$\hat{\mu}_{i}^{\text{MERF}} = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \hat{u}_i = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i}\left(\frac{1}{n_i}\sum_{j \in s_i}(y_{ij} - \hat{f}^{\text{OOB}}(\mathbf{x}_{ij}))\right), \tag{3}$$

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

While the RF part $\hat{f}()$ express the conditional mean of fixed effects, we maintain in Krennmair & Schmid (2022) that $\hat{u}_i$ is the BLUP for the linear part of Model (1). For non-sampled areas, the proposed estimator for the area-level mean reduces to the fixed part from the RF: $\hat{\mu}_i = \bar{\hat{f}}_i(\mathbf{x}_{ij})$.

The access to auxiliary population micro-data is challenging for practitioners, researchers, and even within gatekeeper organizations. The direct incorporation of aggregated auxiliary information in Equation (2) is not possible without misspecification, as for RFs $f(\bar{\mathbf{x}}_i) \neq \bar{f}_i(\mathbf{x}_{ij})$. Notably, 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). Recently, Krennmair et al. (2022) solved this issue by incorporating aggregate census-level covariate information through calibration weights w_{ij}, which balance unit-level predictions from MERFs in Equation (2) achieving coherence with the area-wise covariate means from census data. In short, this estimator under reduced information for the area-level means can be written as:

$$\hat{\mu}_i^{\text{MERFagg}} = \sum_{j=1}^{n_i}\hat{w}_{ij}\left[\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i\right]. \tag{4}$$

However, optimal estimated 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), note that $\mathbf{x}_{ij}$ are unit-level covariates from the survey. Aggregated auxiliary population information ($\bar{\mathbf{x}}_{\text{pop},i}$) is incorporated through optimal weights $\hat{w}_{ij}$ inspired by Li et al. (2019) maximizing the profile empirical likelihood 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 ($\bar{\mathbf{x}}_{\text{pop},i}$) 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 are the solution to the system of equations and obtainable using the Lagrange-multiplier method (Owen, 1990, 2001; Emerson & Owen, 2009). Krennmair et al. (2022) discuss technical conditions for the feasibility of solutions in the context of SAE and propose a best practice strategy, which is compared to predominate methods in model-based SAE as well as the MERF-based estimator under unit-level data from Equation (3).

2.3 Estimation of uncertainty

A discussion on the quality of domain-specific indicators necessitates a scrutiny of inference and uncertainty. For SAE, it is convenient to use the estimated MSE of the indicators. However, even in the supposedly simple case of LMMs with block diagonal covariance matrices and estimated variances, analytical forms of the MSE can only be approximated (Prasad & Rao, 1990; Datta & Lahiri, 2000; González-Manteiga et al., 2008; Rao & Molina, 2015). The deficiency of general statistical results concerning inferences of RFs adds additional complexity. Although, from a survey perspective, Dagdoug et al. (2021) recently analyse theoretical properties of RFs in the context of model-assisted estimation methods, we propose the use of elaborate bootstrap-schemes for the assessment of uncertainty under the previously discussed methods above.

In Krennmair & Schmid (2022), we propose a non-parametric random effect block bootstrap framework for estimating the MSE for area-level means from sampled and unsampled domains as discussed given by Model (3). In short, the bootstrap-schemes builds on non-parametric generation and resampling of random components originally introduced by Chambers & Chandra (2013). Important for handling and resampling the empirical error components is to centre and scale 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), $\hat{\sigma}_\epsilon^2$ is positively biased capturing excess uncertainty concerning the estimation of function $\hat{f}$. We argue a necessity to extrapolate this excess uncertainty before a full bootstrap pseudo-population is simulated. In the presence of aggregated census-level data, Model (4), we base the general procedure on the methodological principles of the bootstrap for finite populations introduced by González-Manteiga et al. (2008). This allows us to construct (pseudo-)true values by generating only error components instead of simulating full bootstrap populations. Details on the methodologies and the performance of proposed uncertainty estimates can found in Krennmair & Schmid (2022) and Krennmair et al. (2022).

3 Illustrative example

This section outlines the advantages of MERFs by estimating domain-level average equivalized household income for Austrian districts. Especially highly skewed distributed variables, like the household income in Austria, often lead to model violations for the classical nested error regression model from Battese et al. (1988). Therefore, semi-parametric methods for SAE, like MERFs, are very promising and needed for these kinds of empirical questions.

The used dataset consists of synthetic Austrian European Union Statistics on Income and Living Conditions (EU-SILC) from 2006 on household-level. Note that this data is exemplary data made publicly available as part of the *R*-package *emdi* (Kreutzmann et al., 2019), which contains detailed information on the data generation process. The major advantage of this illustrative example using

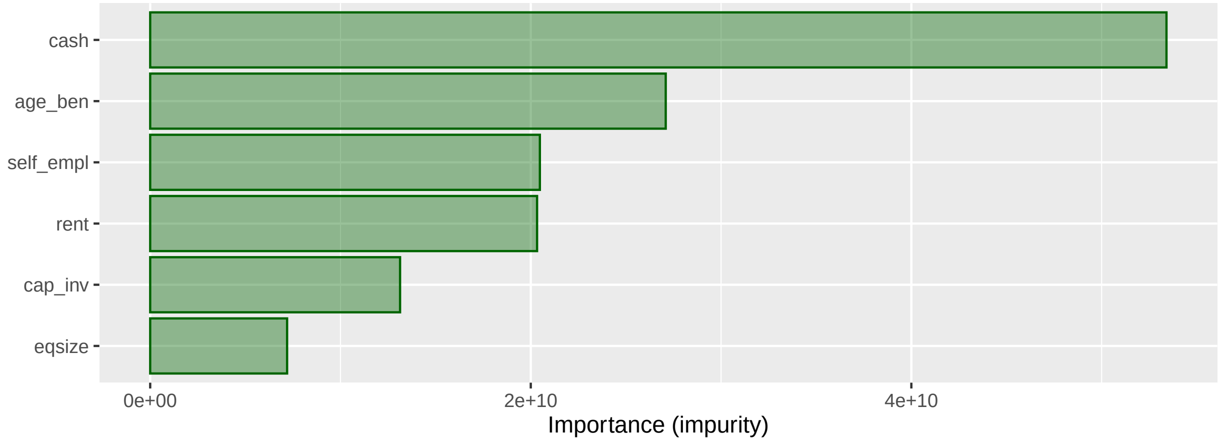

Figure 1: Variable importance for the six most influential variables

open-source data is that we provide reproducible research. The target variable is the equivalized household income (*eqIncome*), which is available in the survey but not in the census and is defined by the ratio of total household disposable income and the equivalized household size (Hagenaars et al., 1994). The illustrative Austrian population data consists of 25 000 households spread over all 94 district and 1945 households within the exemplary sample data. 70 districts are included in the sample, with sample sizes varying between 14 and 200 households (median 22.5 households). Therefore, direct area-level mean-estimates are not feasible for 24 out-of-sample districts. For this reason, and to obtain more precise estimates, SAE methods are needed.

This example displays in addition to the direct estimation, the two MERFs, Model (3) and (4), and the established EBP method (Molina & Rao, 2010) with data-driven Box-Cox transformation (Rojas-Perilla et al., 2020) as competitor. We refer to this method as EBP-BC. Please note that the MERF from Model (3), labelled as MERF_ind, as well as the EBP-BC method require micro-level population auxiliary data. Due to data security constraints, especially in developed countries, alternative estimators relying only on area-level aggregated auxiliary data are highly needed and therefore MERF_agg (from Model (4)) is also included into this example. We aim to show that mean-estimates of MERF_ind are close to the estimates from the established EBP-BC. In addition, the MERF_agg using less data is intended to be similar to both estimators using unit-level auxiliary data.

Regarding variable selection, there is a distinct difference between the EBP method and the MERFs: For EBP-BC, 13 auxiliary variables on socio-economic characteristics and income situation were selected using Bayesian Information Criterion as valid predictors for the target variable *eqIncome*. In contrast, MERFs perform an implicit variable selection. An importance plot gives the reader an impression on most influential variables for the prediction of *eqIncome*: among others, this plot highlights variables describing cash assets (cash), the receiving of age benefits (age_ben), a given situation of self-employment (self_empl) as well as income from rental of a property or land (rent) as particularly influential (cf. Figure 1). Figure 2 shows a line plot on point estimates for all four methods. The direct estimator as well as the EBP-BC are produced using the *R*-package *emdi* (Kreutzmann et al., 2019) and the code for the two MERF estimators is available from authors upon request. The two MERF estimators perform similarly to the established EBP-BC, which confirms their validity. Even under limited population data (MERF_agg), similar results are obtained as with the two methods using micro-level population data. The assessment of uncertainty of point estimates is an important step for an analysis of reliability of estimates. Thus, Figure 3 reports corresponding bootstrap MSE-estimates for point estimates of area-level means. As anticipated, the three model-based estimators are characterized by lower MSE-values in mean and median terms. For MERF_agg, however, this reduction is less pronounced than for the other two estimators, which assume access to comprehensive

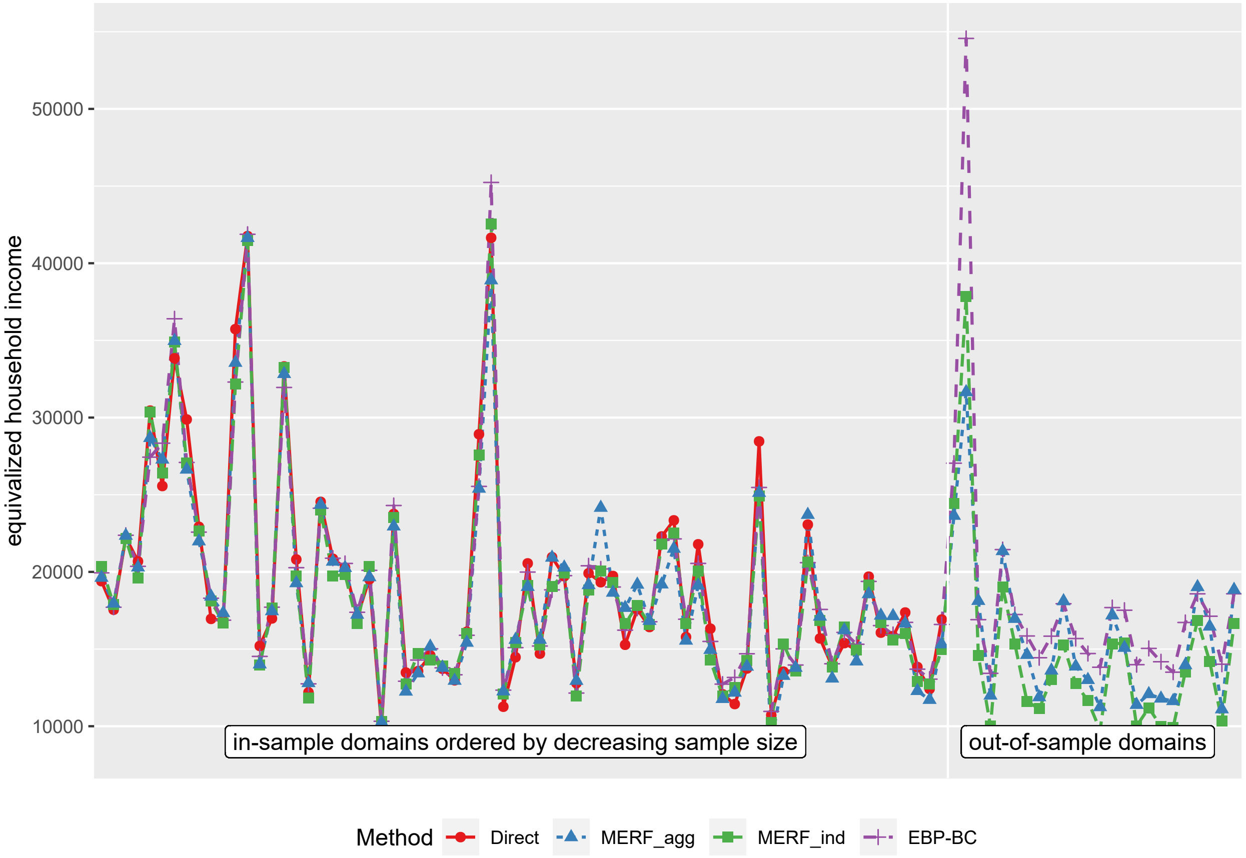

Figure 2: Point estimates for the domain-level average equivalized household income for Austrian districts.

micro-level population data. In median terms, the MSE-values of MERF_ind are the lowest among all competing methods. For detailed model-based simulations and extensive discussions, we refer to Krennmair & Schmid (2022) and Krennmair et al. (2022). The extensive analysis of properties of MERF_ind and MERF_agg reveals that, especially in the presence of complex and unknown relations between covariates, these semi-parametric methods offer substantial advantages.

4 Conclusion and Outlook

Machine learning methods became popular alternatives for predictive models in various scientific fields outside the statistical spheres of SAE. This article serves as a first step, of bridging concepts and highlighting opportunities such as the similarity of the predictive character of model-based SAE and the training/test-set paradigm in machine learning. We introduce RFs for SAE and account for dependency structures of observations using a semi-parametric framework of MERFs for the estimation of point and uncertainty estimates for domain-level indicators under unit- and aggregated auxiliary information. A reproducible example on open-source income data shows estimates for MERFs using unit-level and aggregated auxiliary data and compares them to direct estimates and the well known EBP method (Molina & Rao, 2010) with Box-Cox transformation (Rojas-Perilla et al., 2020). Benefits of RFs are the implicit model-selection and lack of specification under simultaneously high predictive power even in the presence of complex and potentially non-linear interactions between covariates. Moreover, RFs also deal with high-dimensional ($p > n$) datasets. We acknowledge that compared to predominant LMMs, the benefits of prediction serve at cost of explainability and attribution and although this 'black-box'-argument is mitigated by diagnostic tools and plots, discrepancies regarding perspectives of predictive algorithms and explanatory models remain (Efron, 2020).

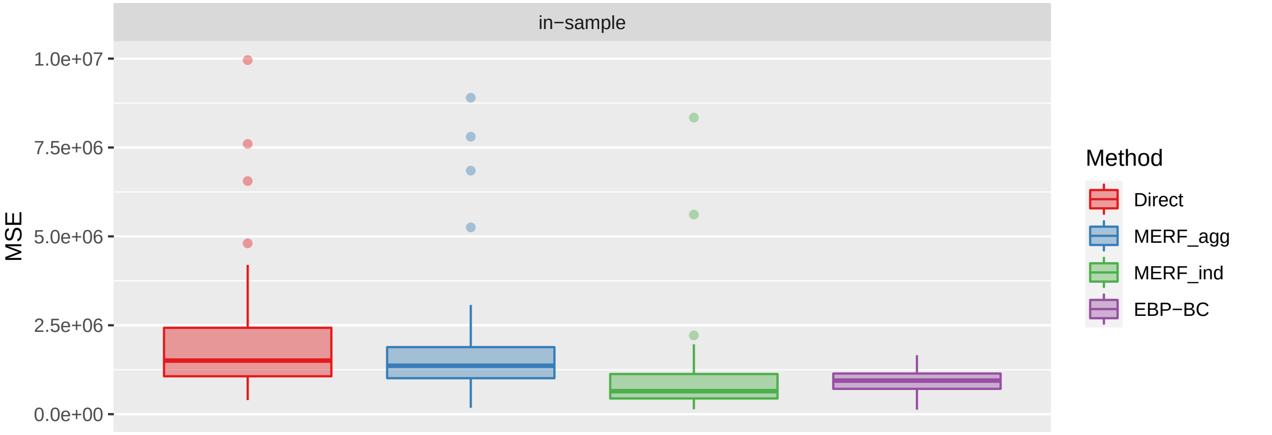

Figure 3: MSE estimates for the point estimates on average equivalized household income for Austrian districts for in-sample areas.

Overall, we conclude that machine learning methods add valuable insights and advantages to the existing repertoire of SAE methods. From our perspective, tree-based predictors perfectly align with the required emphasis on robustification of models against model-failure (e.g. providing insurances against model-misspecification, valid variable selection and the effective handling of outliers) (Jiang & Rao, 2020). The broadening of our statistical methodological toolbox must not only lie in the plain application of existing machine learning algorithms, but rather in the question how they can be made 'scientifically applicable' (Efron, 2020). For SAE, emerging methods need a clear commitment to the methodological tradition of SAE, meaning to find solutions within the context of domain-level indicators, dependent data structures, and in the broader context of survey methodology.

Our presented framework for MERFs, Model (1), is at a starting point and opens up many further research directions. Future applications might use MERFs in the presence of more complex dependency and correlations structures and increasingly compare them to existing LMM-based alternatives. The use of complex and high-dimensional covariate data is another interesting topic. Generally, there is a need for a substantial theoretical discussion on non- or semi-parametric models handling dependency structures. Concretely, our framework can be generalized to binary and count data, but also towards other model-classes, such as Support Vector Machines, Gradient Boosting, Bayesian Additive Regression Trees and many more. We firmly believe that methodological developments in SAE should be complemented by the development of suitable open-source software packages and we are currently working on an *R*-package. Facilitated access to SAE-methods promotes further development and facilitates the comparison between existing methods in model- and design-based evaluations and will result in a toolbox of tailored approaches for researchers and practitioners.

References

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

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.

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

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

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

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

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., & Tzavidis, N. (2006, 06). M-quantile models for small area estimation. *Biometrika*, *93*(2), 255-268.

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.

De Moliner, A., & Goga, C. (2018, December). Sample-based setimation of mean electricity consumption curves for small domains. *Survey Methodology*, *44*(2), 193–215.

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

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

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

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

Gelman, A., & Vehtari, A. (2021). What are the most important statistical ideas of the past 50 years? *Journal of the American Statistical Association*, *116*(536), 2087-2097.

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.

Hagenaars, A. J., De Vos, K., Asghar Zaidi, M., et al. (1994). Poverty statistics in the late 1980s: Research based on micro-data.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-Effects Random Forest for Clustered Data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328.

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

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

Krennmair, P., & Schmid, T. (2022). *Flexible domain prediction using mixed effects random forests.* (Working Paper)

Krennmair, P., Würz, N., & Schmid, T. (2022). *Analysing opportunity cost of care work using mixed effects random forests under aggregated census data.* (Working Paper)

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), 1–33.

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

Marchetti, S., & Tzavidis, N. (2021). Robust estimation of the Theil index and the Gini coeffient for small areas. *Journal of Official Statistics*, *37*(4), 955–979.

Mendez, G. (2008). *Tree-based mehtods to model dependent data* (Unpublished doctoral dissertation). Arizona State University.

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

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

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

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*. New York: Chapman and Hall.

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

Prasad, N. G. N., & Rao, J. N. K. (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., & Molina, I. (2015). *Small area estimation* (2.nd ed.). New Jersey: John Wiley & Sons.

Rojas-Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(1), 121–148.

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

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

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.