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

Nora Würz[*], Timo Schmid[**], and Nikos Tzavidis[***]

[*]Institute of Statistics and Econometrics, Freie Universität Berlin, Berlin, Germany
[**]Institute of Statistics, Otto-Friedrich-Universität Bamberg, Bamberg, Germany
[***]Department of Social Statistics and Demography and Southampton Statistical Sciences Research Institute,
University of Southampton, UK

Abstract

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

Keywords: Census; Density estimation; Official statistics; Unit-level models; Small area estimation

1 Introduction

Reliable knowledge of the spatial distribution of income and wealth is essential for evidence-based policymaking. High spatial resolution direct estimates of income that use household surveys are likely to be unreliable because of the small sample sizes at the spatial scale of interest. A possible way to overcome this problem is by using small area estimation (SAE) methods (Rao and Molina, 2015; Tzavidis *et al.*, 2018). The estimation of income indicators, including non-linear ones, for example the poverty gap and poverty severity, has been researched extensively (Elbers *et al.*, 2003; Molina and Rao, 2010; Tzavidis *et al.*, 2018). These indicators are typically estimated by using model-based methods that assume access to auxiliary information from population micro-data. In developed countries like Germany and the UK, population micro-data are not publicly available and access to such data is even challenging within gatekeeper organizations. Instead, population-level auxiliary data are often only available at some aggregate level.

One predominant approach - for estimating the average income in small areas - is the nested error regression (NER) model proposed by Battese *et al.* (1988) that borrows strength by using auxiliary information from the census. The model relies on the assumption that the error terms follow a Gaussian distribution. In a variety of real-world examples, this assumption may not be satisfied. Especially skewed variables, like income and consumption, can often not be adequately described by the available auxiliary variables and therefore lead to error terms that are not normally distributed. One promising approach to satisfy the assumptions of the NER model is to use fixed logarithmic (Molina and Martín, 2018) or data-driven (Sugasawa and Kubokawa, 2019; Rojas-Perilla *et al.*, 2020) transformations for the dependent variable. Data-driven transformations contain a transformation parameter that adapts to the particular shape of the data. Many publications focus on the estimation of the average income under a logarithmic transformation within the NER model (lognormal model). A general problem is the correction of the bias when the response is back-transformed to the original scale. For the lognormal model, Karlberg (2000) proposes a bias-corrected estimator for the small area mean. Chandra and Chambers (2011) present model-calibrated weights to obtain a bias-correction for a log transformed response variable. Berg and Chandra (2014) suggest an estimator with minimal mean squared error (MSE). Molina and Martín (2018) also focus on this estimator and develop an analytical MSE estimator. However, in all the research work mentioned, auxiliary information from population micro-data is required for the bias-correction due to the back-transformation, which is a strong limitation for data analysts. In contrast, Li *et al.* (2019) propose a bias-correction when only aggregate population-level auxiliary information is available. Their estimator uses the smearing approach of Duan (1983) to build a pseudo-population from the sample data, which is later adjusted by incorporating population means from the auxiliary information. However, Li *et al.* (2019) only discuss results for point estimation and they do not present a MSE estimator. It should also be mentioned that the estimation of small area means/ averages can be also implemented with area-level linear mixed regression models (Fay and Herriot, 1979) when only aggregated data for the survey and the population data are available. For these area-level models, Slud and Maiti (2006) present an estimator for small area means and its analytical MSE estimator under a log transformed Fay-Herriot model. Sugasawa and Kubokawa (2017) discuss area-level models for data-driven dual power transformations.

In this work, we propose methodology for estimating small area means based on the transformed NER model when only aggregate population-level auxiliary information is available. In the absence of population micro-data, appropriate bias-corrections for small area prediction are presented. Under the proposed approach we do not make any parametric assumptions about the auxiliary variables and instead use aggregate statistics (means and covariances) and kernel density estimation (KDE) to resolve the issue of not having access to population micro-data. Regarding the estimation of the MSE, we propose a parametric bootstrap that captures the uncertainty due to the use of transformations and KDE. We study our proposed estimator in extensive model-based and design-based simulations and compare the proposed method to alternative methods for example, the EBP (Molina and Rao, 2010) under transformations (assuming the availability of unit-level census data) and the estimator of Li *et al.* (2019). Results show that, compared to alternative methods, the proposed methodology leads to comparable results. We also show that the proposed uncertainty estimation works.

The proposed methodology is applied to the 2011 Socio-Economic Panel (SOEP) and aggregate census information from the same year to estimate the average individual income for 96 regional planning regions (RPRs) in Germany. Knowledge of the spatial distribution of income in Germany is of great interest: Kosfeld *et al.* (2008) investigate disparities of regional German income at district level for the year 2004. They note that income varies considerably between districts. The districts with the highest and lowest incomes in Germany differed by a factor of 1.8. For Eastern Germany, income is on average 86.3% of that in Western Germany. Within Germany, there is a strong interest in income differentials, particularly between Western and Eastern Germany, which are politically relevant and widely discussed (Frick and Goebel, 2008; Görzig *et al.*, 2008; Fuchs-Schündeln *et al.*, 2010; Kohn and Antonczyk, 2013).

The rest of the paper is structured as follows. Section 2 describes the SOEP survey and discusses initial direct estimators. Furthermore, the auxiliary aggregated information from the census is presented. In Section 3 the methodology is explained. To start with, the classical NER model is reviewed in Subsection 3.1. Afterwards, the transformed NER models and corresponding small area estimators - when population micro-data are available - are discussed in Subsection 3.2. Finally, the proposed method - when only aggregate population-level auxiliary information is available - is introduced in Subsection 3.3. MSE estimation is discussed in Section 4. The proposed methods are evaluated against existing competitors using model-based (Section 5) and design-based simulation studies (Section 6). For the design-based simulation study, individual income data from the Mexican census are used to assess the proposed estimator on real-world data. In Section 7 we apply the proposed method to the SOEP data for estimating average income for German RPRs, where population micro-data are not available. Section 8 summarizes the main findings and outlines further research.

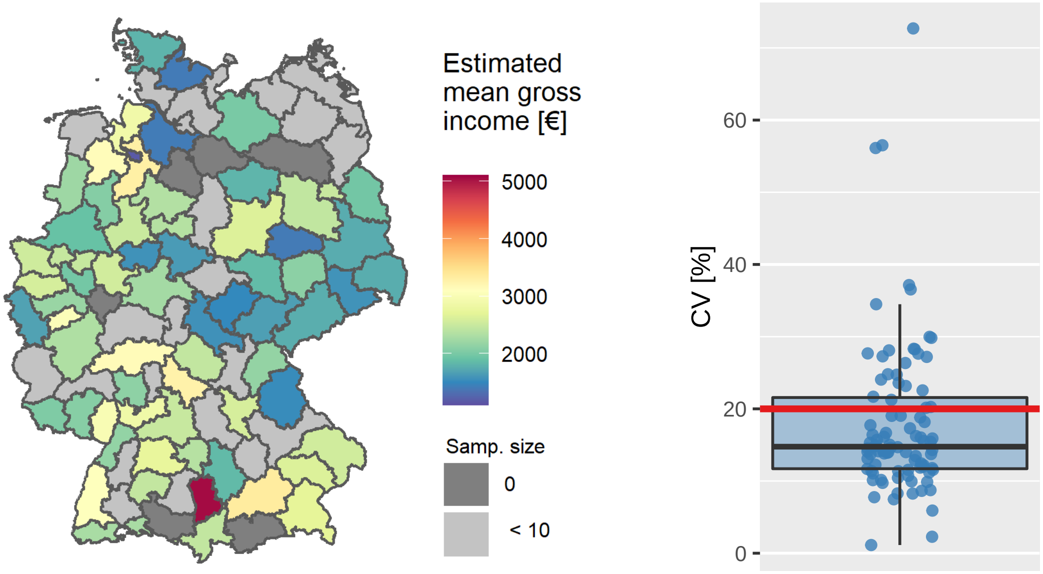
2 Data Sources and Initial Analysis

The current section describes the data sources used in the application. First, the survey *Socio-Economic Panel* is described and the corresponding direct estimates for German RPRs are shown in Subsection 2.1. The direct estimates indicate the need for using SAE methods. In Subsection 2.2, the available auxiliary variables from the German census are described. Finally, this subsection closes with a motivation for the use of transformations in order to meet the model assumptions in the application.

2.1 The German Socio-Economic Panel and initial estimates on spatial gross income

To estimate income in Germany we use the SOEP (Socio-Economic Panel, 2019). This survey was established in 1984, is one of the most important German surveys, and is located at the German Institute for Economic Research (DIW Berlin). The SOEP provides representative longitudinal data of private households in Germany for multidisciplinary issues (Goebel *et al.*, 2019). This sample is highly valuable not only for governmental institutions, but also for researchers from various fields. It collects a large variety of variables and offers different sub-samples with a specific focus. Being interested in estimating the average individual income in Germany for different RPRs, the SOEP is a valuable data source. In other important German surveys, such as the Microcensus, income is provided only in interval-censored groups.

The analysis is conducted with the open-source software R (R Core Team, 2020). We use the refreshment sample from 2011 for our analysis. The sampling design for the 2011 refreshment sample is a multistage stratified sampling procedure. In the first stage, 370 primary sampling units (PSUs) are selected proportionally to their size. For this purpose, stratification into federal states, governmental



(a)

(b)

Figure 1: (a) Map with direct estimated mean gross individual income per month [€] where out-of-sample areas and areas with sample size under ten are greyed out and (b) the respective CVs

regions, and municipalities is carried out. In the second stage, using the random walk methodology, addresses are sampled within each PSU. Within this selection process, migration background was accounted for so that migrant households were twice as common in the 2011 Refreshment sample. The persons living in the respective households are interviewed (Kroh *et al.*, 2018). Our target variable is the gross individual income in Euro of the respondent in the month before the interview during the year 2011. We choose this year because the census was carried out in the same year and so this choice enables us later to include census covariates. The target population is the working age population, which is defined as those aged 15 to 64 (OECD, 2020). We are interested in estimating the mean gross individual income in 96 German RPRs. Figure 1 (a) shows the direct estimates, calculated with the emdi-package (Kreutzmann *et al.*, 2019). The mean gross individual income varies from 1173 € (Bremen) to 5059 € (Donau-Iller). General trends can be seen from the map: East Germany has a lower mean gross individual income, the Munich, Stuttgart, or Frankfurt areas have higher mean incomes. However, small sample sizes lead to estimated mean gross individual incomes with high variances. With dark grey we present the six out-of-sample regions. Furthermore, for 23 regions, the sample size is less than ten (presented in light grey). In these cases, we are not allowed to report directly estimated mean gross individual incomes due to confidentiality agreements with the data provider. Therefore, direct estimates are presented for 67 out of 96 regions. The sample sizes over the RPRs vary from 0 to 107 (1st Quartile: 8, Median: 16.5, Mean: 20.83, and 3rd Quartile: 26.5). We use a calibrated bootstrap method (Alfons and Templ, 2013) which accounts for the survey design to estimate the variance of the direct estimates implemented in the R package emdi (Kreutzmann *et al.*, 2019). From these variances the corresponding coefficients of variation (CV) are obtained (cf. Figure 1 (b)). In particular, Eurostat considers estimators with a CV less than 20% to be reliable (Eurostat, 2019). Here, 26 out of 90 CVs exceed this threshold for reliable estimates. The highest CVs were found for Osthessen (72.7%), Bayerischer Untermain (56.5%), and Donau-Iller (56.13%). When looking at the individual sample data, on the one hand the sample sizes are small for these RPRs and on the other hand the variability of the reported individual incomes is high. Therefore, exploring the use of SAE may be necessary if interest is in improving the estimation accuracy for RPRs. Since some of the SOEP auxiliary variables are measured in the same way within the German census, census covariate data can serve as auxiliary information in small area models. However, information from the German census is only available at aggregated (RPR) level.

2.2 Auxiliary data from the German census and preliminary model selection

SAE methods use survey data and population-level auxiliary information to improve the available direct estimates. Especially for small sample sizes, as in the 2011 SOEP data disaggregated into the 96 German RPRs, the methods might be helpful in order to achieve more reliable estimates. As the German census is not available at population micro-level, we use aggregated auxiliary information (means and covariances) from the German census 2011 (Statistisches Bundesamt, 2015) to estimate the mean gross individual income. We first made a pre-selection of suitable variables that are available in the SOEP and in the German census. Subsequently, we use model selection criteria (the conditional Akaike information criterion from the cAIC4-package (Saefken *et al.*, 2021)) to select the best model based on the survey data (SOEP) at individual-level. Due to the sampling process, migration background was included by default as a variable in the model. The additionally selected five variables are sex, age, position in the household, employment status and tenant or owner of dwelling. Table 5 in the appendix provides further information about each variable. In order to implement the proposed methodology, we requested the corresponding means and covariances of the six variables from the 2011 German census from the Statistical Office (DESTATIS).

For skewed variables, like gross individual income in our case, transforming the response is often necessary and offers a promising approach to satisfying the assumptions of the commonly used models (Rojas-Perilla *et al.*, 2020). Therefore, NER models have been constructed using the sample data, whereby the dependent variable was on the original scale, log, and log-shift transformed scales. The log-shift transformation is a data-driven transformation which extends the log transformation by a shift-parameter λ: $\log(y + \lambda)$ estimated using the data. For more information about this transformation, see Subsection 3.2. The validity of the normality assumptions for both error terms (unit-level and area-level) of the underlying NER models are checked with QQ-plots (cf. Figure 2). If the original scale or the log transformation is used, deviations from the normal distribution can be seen for the residuals at both levels. Furthermore, we compute the conditional coefficient of determination (Nakagawa and Schielzeth, 2013): 11.99% (no transformation), 36.01% (log transformation), and 41.98% (log-shift transformation). From these preliminary investigations, the decision to use a log-shift transformation for estimating mean gross individual income for German RPRs becomes clear. In Table 5 in the appendix, the coefficients for the chosen NER model with log-shift transformed response (with estimated shift-parameter: $\hat{\lambda} = 358.73$) are shown. Furthermore, the likelihood ratio test leads to clear significance (p = 7.02310^{-15}) of the random effect.

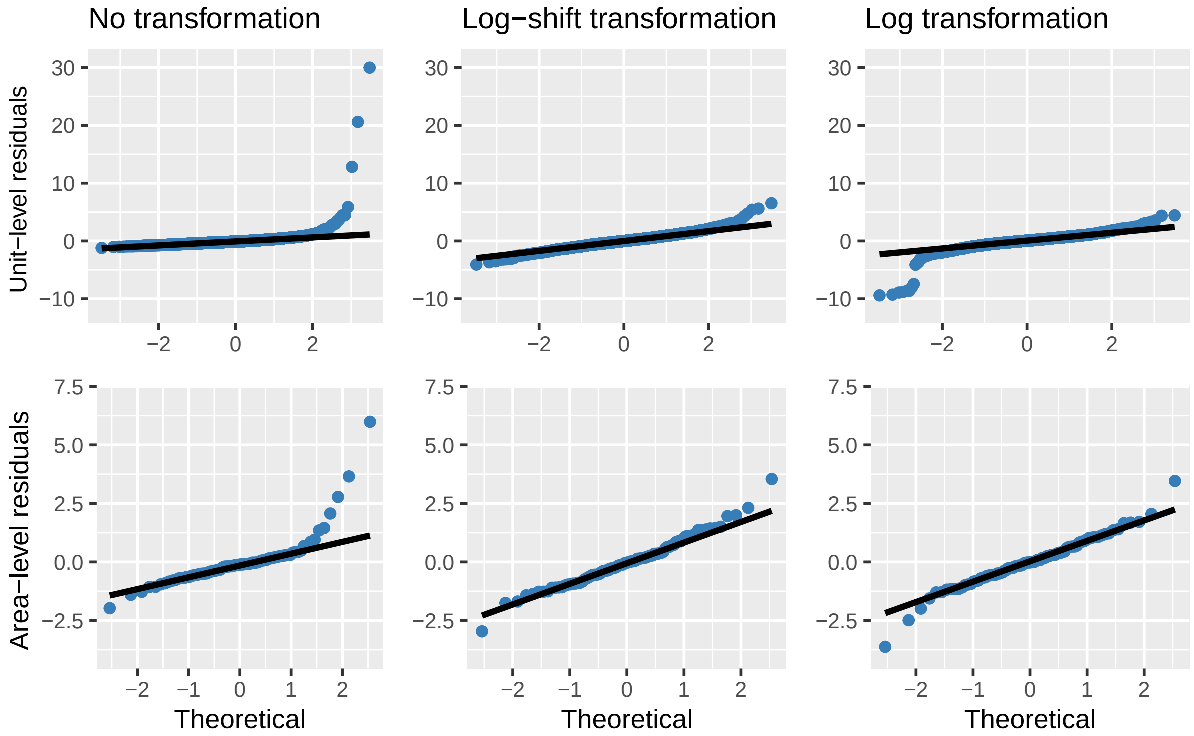


Figure 2: QQ-plots for the individual-level residuals (1st row) and the area-level residuals (2nd row) among the NER models under different transformations

3 Unit-level small area models

The structure of this section is as follows. First, the NER model of Battese *et al.* (1988) is presented in Subsection 3.1. Subsequently, the NER model is extended to allow the use of transformations on the dependent variable (Berg and Chandra, 2014; Molina and Martín, 2018; Rojas-Perilla *et al.*, 2020). Computing the small area estimators requires access to population micro-data to avoid introducing back-transformation bias. As a result, we propose an estimator which can be used when the response is transformed and we have access only to limited auxiliary information in the form of population-level aggregates of the covariates.

3.1 The nested error regression model

The finite population U of size N is divided into D areas $U_1, U_2, ..., U_D$ consisting of $N_1, N_2, ..., N_D$ units. The index $i = 1, ..., D$ indicates the respective area and $j = 1, ..., N_i$ the corresponding units. The continuous dependent variable y_{ij} is available for every unit in the sample s. The sample s consists of n units partitioned into sample sizes $n_1, n_2, ..., n_D$ for the particular areas. With s_i we refer to the in-sample units in area i and with $\overline{s}_i$ to the $N_i - n_i$ non-sampled units in area i. Furthermore, a vector $\mathbf{x}_{ij} = (1, x_1, x_2, ..., x_p)^T$ consisting of the intercept and p explanatory variables is available for every unit j in every area i within the sample. The matrix X_s contains all covariates of every individual in the sample across all areas. Battese *et al.* (1988) use a NER model, which models the relationship between $\mathbf{x}_{ij}$ and y_{ij} as follows:

$$y_{ij} = \mathbf{x}_{ij}^T\beta + u_i + e_{ij}, \qquad u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2) \text{ and } e_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2), \tag{1}$$

where $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ is the vector of regression coefficients, which describes the linear relationship of $\mathbf{x}_{ij}$ and y_{ij}. u_i denotes the area-specific random effect and e_{ij} is the unit-level error. The error terms u_i and e_{ij} are assumed to be independent and σ_u^2 and σ_e^2 denote the corresponding variances. The best linear unbiased predictor (BLUP) for every out-of-sample unit $j \in \overline{s}_i$ is given by

$$\mu_{ij} = \mathbf{x}_{ij}^T\beta + u_i = \mathbf{x}_{ij}^T\beta + \gamma_i \left(\sum_{j \in s_i} \left(y_{ij} - \mathbf{x}_{ij}^T\beta \right) \right), \tag{2}$$

where $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2/n_i}$ denotes the shrinkage factor. The target parameter is the population mean for each area i. The empirical best linear unbiased predictor (EBLUP) for the population area mean ($\overline{y}_i$) is defined as

$$\begin{aligned} \hat{\overline{Y}}_i^{\text{BHF}} &= \frac{1}{N_i} \left(\sum\nolimits_{j \in s_i} y_{ij} + \sum\nolimits_{j \in \overline{s}_i} \hat{\mu}_{ij} \right) \\ &= \hat{\gamma}_i \left(\frac{1}{n_i} \sum_{j \in s_i} y_{ij} + \left(\overline{\mathbf{x}}_i - \frac{1}{n_i} \sum_{j \in s_i} \mathbf{x}_{ij} \right)^T \hat{\beta} \right) + (1 - \hat{\gamma}_i) \, \overline{\mathbf{x}}_i^T \hat{\beta}, \end{aligned} \tag{3}$$

where $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i}$. The vector $\overline{\mathbf{x}}_i^T = \frac{1}{N_i} \sum_{j \in U_i} \mathbf{x}_{ij}^T$ denotes the p population means for each area i. For estimating fixed effects and the variance components σ_u^2 and σ_e^2, several methods are available, for example maximum likelihood (ML) or restricted maximum likelihood (REML) (Rao and Molina, 2015). Note that the estimator in (3) requires access only to population-level aggregates ($\overline{\mathbf{x}}_i^T$) and to unit-level survey data.

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

In a large number of applications, the dependent variable has a skewed distribution, and its shape cannot be sufficiently explained by the available auxiliary variables. Consequently, the Gaussian assumptions for the random effects and the error terms in the NER model might be violated. One common solution to address this problem is to use one-to-one transformations $h(y_{ij}) = y_{ij}^*$. In many applications related to income, a fixed logarithmic transformation is used for this purpose due to its simplicity. More flexible solutions depending on the particular shape of the dependent variable are offered by data-driven transformations (Gurka *et al.*, 2006; Rojas-Perilla *et al.*, 2020). For instance, the log-shift transformation (Yang, 1995) extends the log transformation by including a transformation parameter λ: $y_{ij}^* = h(y_{ij}) = \log(y_{ij} + \lambda)$, which is estimated from the sample. For further details we refer the reader to Rojas-Perilla *et al.* (2020).

The use of a transformation for the dependent variable defines a model on the transformed scale (transformed NER model):

$$h(y_{ij}) = y_{ij}^* = \mathbf{x}_{ij}^T\beta + u_i + e_{ij}, \qquad u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2) \text{ and } e_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2). \tag{4}$$

The BLUP on the transformed scale for out-of-sample units is analogous to (2) $\mu_{ij}^* = \mathbf{x}_{ij}^T\beta + u_i$. In most applications, the goal is to estimate the target parameter (here the mean) on the original scale. From Jensen's inequality, it is obvious that the naive back-transformation for the BLUP of the out-of-sample predictions doesn't lead to the same result as the best prediction on the original scale. If the transformation $h()$ is a real convex or concave function the following applies because of Jensen's inequality (Jensen *et al.*, 1906)

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

In this paper, we focus on the log and log-shift transformation, so the back-transformation $h^{-1}() = \exp()$ or $h^{-1}() = \exp() - \lambda$ is in both cases convex. Consequently, $\mu_{ij}^{\text{trans, naive}}$ leads to lower estimates than $E[h^{-1}(y_{ij}^{*})|\mathbf{y}_s, \mathbf{X}_s]$ and thus the estimated small area means using the naive back-transformation

$$\hat{\overline{Y}}_i^{\text{trans, naive}} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} \hat{\mu}_{ij}^{\text{trans, naive}}\right) = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} h^{-1}\left(\mathbf{x}_{ij}^T\hat{\beta} + \hat{u}_i\right)\right) \tag{5}$$

are smaller than the small area means constructed from $E[h^{-1}(y_{ij}^{*})|\mathbf{y}_s, \mathbf{X}_s]$. In the case of a log-transformation, Berg and Chandra (2014) and Molina and Martín (2018) propose an analytical bias-correction. The best predictor for the out-of-sample units is defined for general transformations via an integral which can be solved analytically for $h() = log()$ by using $y_{ij}^{*}|\mathbf{y}_s, \mathbf{X}_s \sim \mathcal{N}\left(\mu_{ij}^{*}, \sigma_u^2(1-\gamma_i) + \sigma_e^2\right)$ - with corresponding density $f_{y_{ij}^{*}|\mathbf{y}_s,\mathbf{X}_s}$ - which comes directly from model (4),

$$\mu_{ij}^{\text{trans, bc}} = E[h^{-1}(y_{ij}^{*})|\mathbf{y}_s, \mathbf{X}_s] = \int_{-\infty}^{+\infty} h^{-1}(y) f_{y_{ij}^{*}|\mathbf{y}_s,\mathbf{X}_s}(y)dy$$

and for the model under log-transformation with $h^{-1}() = \exp()$ applying the property for the expected value of an exponential transformed normally distributed random variable $(E[\exp(X)] = \exp(E[X] + 0.5Var[X]))$ we get

$$\mu_{ij}^{\text{trans, bc}} = \exp\left(\mu_{ij}^{*} + \underbrace{\frac{\sigma_u^2(1-\gamma_i) + \sigma_e^2}{2}}_{=\alpha_\mathbf{i}\text{ (bias-correction)}}\right).$$

To the BLUP (2) on the transformed scale μ_{ij}^{*} a bias-correction α_i is added before applying the back-transformation. By using $\mu_{ij}^{\text{trans, bc}}$ and the estimated coefficients from the sample we obtain the bias-corrected estimator of the small area mean by

$$\hat{\overline{Y}}_i^{\text{trans, bc}} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} \hat{\mu}_{ij}^{\text{trans, bc}}\right) = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} \exp\left(\mathbf{x}_{ij}^T\hat{\beta} + \hat{u}_i + \hat{\alpha}_i\right)\right). \tag{6}$$

As can be seen in (6) for the log-transformation, out-of-sample population micro-data are required for computing the estimator. Again, due to the Jensen's inequality a (second-order) bias is introduced if we use a naive back-transformation of the synthetic part $\exp\left(\overline{\mathbf{x}}_i^T\hat{\beta}\right)$ instead of $\sum_{j\in\overline{s}_i}\exp\left(\mathbf{x}_{ij}^T\hat{\beta}\right)$. The estimator with first-order bias-correction (α_i) and naive back-transformation of the population-level aggregates, is denoted by

$$\hat{\overline{Y}}_i^{\text{bc-naive-agg}} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} \exp\left(\overline{\mathbf{x}}_i^T\hat{\beta} + \hat{u}_i + \hat{\alpha}_i\right)\right). \tag{7}$$

Due to the use of aggregated auxiliary data this estimator has a second-order bias. If we exactly know the underlying distribution of the covariates (for example covariates coming from a normal distribution) it is possible to add an analytical second-order bias-correction, which corrects for the back-transformation of the covariates. Otherwise, alternative solutions to this problem are needed. The assumptions of having access to population micro-data or knowing exactly the joint distribution of the covariates are both strong assumptions in real applications. Thus, the research question that motivates our work is how to estimate small area means under the transformed NER model when only aggregated population-level auxiliary information is available.

3.3 Small area means under limited auxiliary information

Due to the use of a transformation, a bias-correction for the back-transformed estimator of the small area mean is necessary as shown above. This back-transformed bias-corrected estimator (6) requires population-level auxiliary data. In the absence of population micro-data, a second-order bias is introduced if the aggregated covariates are used instead of individual data (7). Our proposed method aims to reduce - additionally to the first-order bias-correction - the second-order bias due to the back-transformation of the synthetic part. Therefore, it offers a solution to deal with bias under limited auxiliary information. The aim of this subsection is to propose an approximation of $\mathbf{x}_{ij}^T\hat{\beta}$ in the absence of population micro-data which is needed to deal with the second-order bias and combining this with the first-order bias-correction (α_i) for small area means.

Kernel-density estimation for the synthetic part: As mentioned above in the presence of limited auxiliary information we are not able to obtain $\left(\sum_{j\in\overline{s}_i}\exp\left(\mathbf{x}_{ij}^T\hat{\beta}\right)\right)$ necessary for computing (6). Therefore, we have to rely on estimation methods for the unknown synthetic part ($\mathbf{x}_{ij}^T\hat{\beta}$) under limited auxiliary information. We propose the use of a KDE approach to estimate the distribution of the synthetic part. We use KDE based on the synthetic part of the model (instead of the auxiliary variables), because the estimated distribution of the synthetic part is sufficient to deal with the second-order bias. We decide to take this approach for the following reasons: First, we can use univariate KDE for the synthetic part $\left(\mathbf{x}_{ij}^T\hat{\beta}\right)$ compared to multivariate KDE which is needed for estimating the joint multivariate distribution of the auxiliary variables. To the best of our knowledge, implementations of multivariate KDEs are only possible with the available packages in R for a maximum number of six auxiliary variables. Especially if categorical variables are used in the model, this limit presents a significant restriction in many applications. Second, no parametric assumptions about the covariates are necessary and only aggregate auxiliary information at the population-level is required, while availability of unit-level sample data is still assumed. Please note, in case of exclusively categorical covariates, the established EBP-method (Molina and Rao, 2010) under transformations (Rojas-Perilla *et al.*, 2020) is suitable after expanding the counts from the cross-tables. The introduced KDE-based method is only appropriate for metric variables or a mixture of metric and categorical variables.

KDE is one of the most popular non-parametric density estimation techniques first mentioned by Rosenblatt (1956) and Parzen (1962). For a general overview, we refer to Scott (2015). Given a sample $X = \{X_1, ..., X_n\}$ KDE estimates the density f by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} k \left(\frac{x - X_i}{h} \right), \tag{8}$$

where the function $k()$ is the kernel and h is the bandwidth. Here we opt for the use of the Epanechnikov kernel (Epanechnikov, 1969), which is implemented using the density-function of stats-package (R Core Team, 2020). Furthermore, for bandwidth selection we use the method from Sheather and Jones (1991), which is widely recommended (Venables and Ripley, 2002). However, other bandwidth selectors can be also applied.

The first step in computing the estimator involves standardising the predictions using the synthetic part of the NER model, which will be later adjusted by using population-level auxiliary information. The standardised predicted values z_{ij} for area i and unit j are computed as follows,

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

For this purpose, the standardisation step uses the mean and the standard derivation from the predictions of the synthetic part of the model.

In a second step, we adjust the predictions with the aggregated population-level auxiliary data: The mean $\overline{\mathbf{x}}_i^T \hat{\beta}$ and the empirical variation $\sigma_{i, \mathbf{X}^T \hat{\beta}} = \sqrt{\sum_{k=0}^{p} \sum_{l=0}^{p} \hat{\beta}_k \hat{\beta}_l \text{Cov}[\mathbf{x}_{ik}, \mathbf{x}_{il}]}$, where $\text{Cov}[\mathbf{x}_{ik}, \mathbf{x}_{il}]$ is the known covariance between the k-th and l-th explanatory variable for area i. In addition to the requirements of the NER model, we assume that these covariances are also known or obtained from the data owner. Since they are area-level values, usually no confidentiality issues arise for statistical offices. This step adds the small area idea to the proposed method by using the aggregated census means and covariances to adjust the standardised predicted values. Since in many small area applications sample sizes considerably differ among the areas, we propose the following distinction. For large sample sizes, we use the standardized data (z_{ij}) from the respective area i (conditional). In contrast, for small sample sizes, we use the standardized data (z_{ij}) from all areas (unconditional). To distinguish between small and large sample sizes, we define a threshold t. If the respective area has a small sample size below the threshold ($n_i < t$) - or is even an out-of-sample area - we use the standardized data from all areas to generate adjusted data for area i. The following formula illustrates how the input values for the KDE (r_{im}) are obtained from the standardised values z_m. The index m ranges from $1, ..., n$ for sample sizes below t (unconditional) and from $1, ..., n_i$ for sample sizes above t (conditional). Note that depending on the sample size within the actual area i and the chosen threshold t a different number of standardised values are used to determine the input values. With

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

as input we estimate the respective density using the KDE (8) for each area i. We denote the resulting density for area i by $\hat{f}_{h,i}$.

Small area means under limited auxiliary information: If the naive back-transformed estimator (5) is used under limited auxiliary information with a naive back-transformation of the synthetic part, we would obtain

$$\hat{\overline{Y}}_i^{\text{trans, naive-agg}} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} h^{-1}\left(\overline{\mathbf{x}}_i^T\hat{\beta} + \hat{u}_i\right)\right). \tag{9}$$

This estimator does not correct for the first- and second-order bias due to the use of a transformation. In order to account for both types of bias the proposed method relies on the approximated area-specific density $\hat{f}_{h,i}$ of the synthetic part and the first-order bias-correction α_i. Starting with (6) we get

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

where $\hat{\mu}_{ij} = \mathbf{x}_{ij}^T\hat{\beta} \; + \hat{u}_i$ is defined analogously to (2). Under limited auxiliary information, the problem is reduced to determining the unknown back-transformed total (T_i). We use numerical integration and the estimated density of the synthetic part $\hat{f}_{h,i}$ to estimate this total $\hat{T}_i$ from the available sample data and the population-level auxiliary information - without using population micro-data. In detail, the total can be expressed as $\hat{T}_i = \sum_{j=1}^{N_i}\exp\left(\mathbf{x}_{ij}^T\hat{\beta}\right) = N_i E[\exp(\mathbf{x}_{ij}^T\hat{\beta})] = N_i\int_{-\infty}^{+\infty}\exp(x)\hat{f}_{h,i}(x)dx$. This integral can be determined by using numerical integration. By inserting the estimated back-transformed area-specific totals $\hat{T}_i$, we obtain the small area estimator of the mean when access to population-level auxiliary information is limited to population-level aggregates

$$\hat{\overline{Y}}_i^{\text{trans, bc-agg}} = \frac{1}{N_i}\hat{T}_i\exp\left(\hat{u}_i + \hat{\alpha}_i\right). \tag{10}$$

For the log-shift transformation, the only change is the use of the data-driven shift-parameter $\hat{\lambda}$ resulting in

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

In Sections 5 and 6 the properties of the proposed estimator are investigated in model-based and design-based simulation studies and compared with other competitors.

4 Uncertainty estimation

Quantifying the uncertainty of small area estimates can be challenging. For the log-transformed NER model, Molina and Martín (2018) derive an analytic MSE estimator and also present a parametric bootstrap MSE estimator that follows the ideas in González-Manteiga *et al.* (2008). Rojas-Perilla *et al.* (2020) propose two bootstrap schemes for estimating the MSE under data-driven transformations. However, both papers assume access to population micro-data. For the proposed estimator (10) when only population-level aggregates are available, we develop a parametric bootstrap MSE that captures the additional uncertainty due to KDE and the estimation of the adaptive shift parameter in the case of a log-shift transformation. The steps of the parametric bootstrap are as follows.

1. Transform the data: $y_{ij}^{*} = h(y_{ij})$, where h is the log or the log-shift transformation.
2. Estimate $\hat{\beta}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ using the transformed NER model (4) and the sample data. In the case of the log-shift transformation, estimate $\hat{\lambda}$ from the sample data as proposed by Rojas-Perilla *et al.* (2020).
3. For $b = 1, ..., B$
   (a) Generate $u_i^{(b)} \sim \mathcal{N}(0, \hat{\sigma}_u^2)$ and $e_{ij}^{(b)} \sim \mathcal{N}(0, \hat{\sigma}_e^2)$ for all areas i and $j \in s_i$.
   (b) Generate bootstrap samples for all areas i on the transformed scale:

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

   and for each bootstrap replication b estimate the small area mean using the proposed estimator $\hat{\overline{Y}}_i^{\text{trans, bc-agg, (b)}}$ (10). If the adaptive log-shift transformation is used, λ is re-estimated for each bootstrap replication b.

   (c) Determine the true mean for each area i in each bootstrap replication b. Note that we cannot reconstruct a bootstrap population because population micro-data for x are not available. However, from the available aggregated population-level values (area-specific means and covariances for the auxiliary information), which are kept constant across the bootstrap replications, we can construct an area-specific distribution on the transformed scale for each bootstrap replication b. The distribution is defined as

$$y_{ij}^{*(b)} | \mathbf{y}_s^{(b)}, \mathbf{X}_s, u_i^{(b)} \sim \mathcal{N}\left(\overline{\mathbf{x}}_i^T \hat{\beta} + u_i^{(b)}, \sigma_{i,\mathbf{X}^T\hat{\beta}}^2 + \hat{\sigma}_e^2\right), \tag{11}$$

   where $\sigma_{i,\mathbf{X}^T\hat{\beta}} = \sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p} \hat{\beta}_k \hat{\beta}_l \text{Cov}[\mathbf{x}_{ik}, \mathbf{x}_{il}]}$ is determined from known covariances and estimated regression coefficients and $\hat{\sigma}_e^2$ is estimated in step 2. For the derivation of the distribution (11), see the additional information in the appendix. Since the aim is to identify the true mean ($\overline{Y}_i^{(b)}$) on the original scale, we need to combine the distributional assumptions on the transformed scale (11) with the properties of the back-transformation function $h^{-1}() = exp()$. Because the area-specific distribution $y_{ij}^{*(b)}$ for each replication b is assumed to be normally distributed with known mean and variance, we can apply the property for the expected value of an exponential transformed normally distributed random variable ($E[\exp(X)] = \exp(E[X] + 0.5Var[X])$). Therefore, we obtain

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

If we use the data-driven log-shift transformation, the analogous equation is

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

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

4. Determine the MSE over the B bootstrap replications:

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

5 Model-based simulation study

In this section, we present results from a model-based simulation study. We evaluate point estimators and the parametric bootstrap MSE estimator under the four scenarios presented in Table 1. For each of the four scenarios we generate finite populations U of size $N = 50000$. The population U is partitioned into 50 areas $U_1, U_2, ..., U_D$ consisting of $N_i = 1000$ units. From each population, we draw a sample by stratified random sampling, where the strata are defined by the 50 areas. The area-specific sample sizes n_i vary between 2 and 58 with median sample size equal to 33. We chose these sample sizes for three reasons: Firstly, we are interested in evaluating the estimators under a scenario where some areas have very small sample sizes. Secondly, we are interested in having areas with a sample size exceeding the threshold of $t = 40$ for using the (conditional) small area KDE estimator. In this case, 19 out of 50 areas exceed this threshold in each drawn sample. Thirdly, the sample sizes are similar to those in the real data application.

The scenarios are labelled *Normal*, *Log-Scale*, *Log-Scale Gamma*, and *GB2* (Generalised Beta distribution of the second kind). Each scenario is repeated independently $M = 500$ times. In the case of the *Normal* scenario, the generating model is linear and both error terms (u_i and e_{ij}) follow a normal distribution. Therefore, it is a reference scenario where no transformation is required. In contrast, the *Log-Scale* scenario represents a typical situation related to income data where the dependent variable follows a lognormal distribution such that a fixed log transformation is required. Under the third scenario (*Log-Scale Gamma*) the auxiliary variable is not normally distributed. We chose this scenario in order to assess the ability of the proposed method to handle non-normally distributed auxiliary variables. The *GB2* scenario provides a more realistic scenario for income. In this case, the unit-level error terms are simulated by using a *GB2* distribution and the random effects are generated by using a normal distribution. Under the *GB2* scenario the use of a data-driven transformation should be more suitable than the use of the untransformed model or of the fixed log transformation.

5.1 Evaluation of the estimated back-transformed totals

In this subsection we investigate the behaviour of the estimated back-transformed totals, before exploring small area means. We compare the back-transformed totals estimated using population micro-data $\left(T_i = \sum_{j=1}^{N_i}\exp(\mathbf{x}_{ij}^T\hat{\beta})\right)$ as in the bias-corrected estimator (*bc*) (6), the proposed KDE-version of $\hat{T}_i$ as in the proposed estimator using population-level aggregates (*bc-agg*) (10), and the naive back-transformed totals $\left(\exp\left(\overline{\mathbf{x}}_i^T\hat{\beta}\right)\right)$ as in the two types of naive back-transformed estimators using popu-

Table 1: Model-based simulation scenarios

Scenario	Normal	Log-Scale	Log-Scale Gamma	GB2
Model	$4500 - 400\mathbf{x}_{ij} + u_i + e_{ij}$	$\exp(10 - \mathbf{x}_{ij} - 0.5\mathbf{z}_{ij} + u_i + e_{ij})$	$\exp(10 - \mathbf{x}_{ij} - 0.5\mathbf{z}_{ij} + u_i + e_{ij})$	$8000-400\mathbf{x}_{ij}+u_i+e_{ij}-\bar{e}$
$\mathbf{x}_{ij}$	$N(\mu_i, 9)$	$N(\mu_i, 0.25)$	$\Gamma(\mu_i, 2) + 1$	$N(\mu_i, 25)$
$\mathbf{z}_{ij}$	-	$N(0, 1)$	$N(0, 1)$	-
μ_i	$U[2, 3]$	$U[2, 3]$	$U[0.8, 1.6]$	$U[-1, 1]$
u_i	$N(0, 500^2)$	$N(0, 0.16)$	$N(0, 0.16)$	$N(0, 500^2)$
e_{ij}	$N(0, 1000^2)$	$N(0, 0.8)$	$N(0, 0.3)$	$\text{GB2}(2.5, 1700, 18, 1.46)$

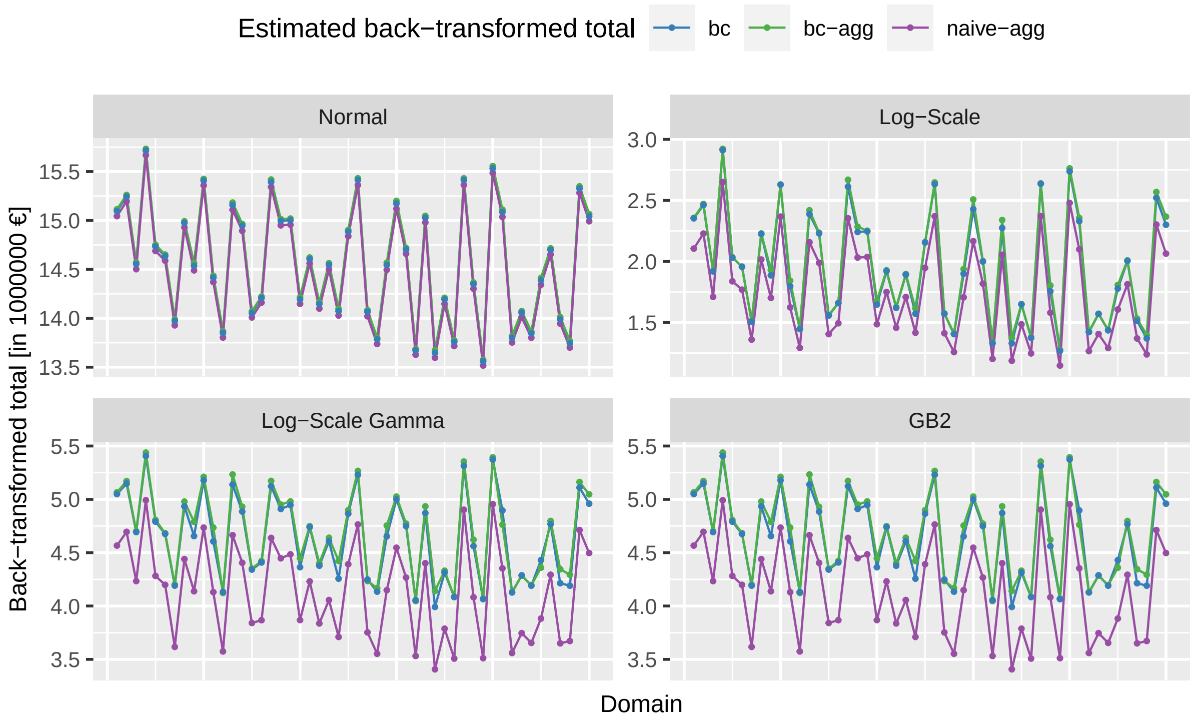


Figure 3: Estimated back-transformed totals *bc* $\left(T_i = \sum_{j=1}^{N_i} \exp(\mathbf{x}_{ij}^T\hat{\beta})\right)$, *bc-agg* $\left(\hat{T}_i\right)$, and *naive-agg* $\left(\exp\left(\bar{\mathbf{x}}_i^T\hat{\beta}\right)\right)$ over the domains for one arbitrarily chosen Monte-Carlo simulation

lation-level aggregates (*bc-naive-agg* and *naive-agg*) (7) and (9).

Figure 3 shows the back-transformed totals for one arbitrarily chosen Monte-Carlo replication under the log-shift transformation. For all four scenarios, the estimated back-transformed totals using population micro-data (*bc*) and the KDE-version (*bc-agg*) show the same behaviour. In the *Log-Scale*, the *Log-Scale Gamma* and the *GB2* scenarios distinctly lower estimates of the domain totals are derived by using the naive back-transformation and this is because of Jensen's inequality. These results underline the importance of the second-order bias-correction if population micro-data are not available and therefore the need of the proposed KDE for estimating the synthetic part.

5.2 Performance of point estimators of the small area means

We compare the proposed estimator (10), denoted *bc-agg*, with existing SAE methods. This includes the direct estimator (*direct*) and the estimator of Battese *et al.* (1988) under the NER model using the untransformed data (*BHF*) implemented in the R package sae (Molina and Marhuenda, 2015). Furthermore, we use different estimators under the log and log-shift transformations. In particular, we compare the proposed bias-corrected estimator to the *naive* back-transformed estimator using population micro-data (5) and two types of naive back-transformed estimator under limited auxiliary information one with no bias-correction (9), denoted *naive-agg*, and one with a first-order bias-correction (7), denoted *bc-naive-agg*. An additional competitor that also aims to tackle the issue of not having access to population micro-data is the *TNER2* estimator of Li *et al.* (2019) which requires only aggregated population-level data. The *TNER2* estimator uses the smearing approach of Duan (1983) to build a pseudo-population, which is then back-transformed to the original scale and finally adjusted with the aggregated population means. For the *TNER2* estimator no MSE estimator has been proposed. Finally, we compare the proposed estimator to the EBP of Molina and Rao (2010) (*EBP*). The EBP is implemented by fitting the NER model under the different simulation scenarios using the original data, a log transformation and an adaptive log shift transformation (Rojas-Perilla *et al.*, 2020). The EBPs are estimated using $L = 100$ Monte-Carlo replications following the recommendations of Molina and Rao (2010). Because the EBPs use auxiliary information from population micro-data they can be treated as a gold standard.

To compare the different estimators, we compute the relative bias and root mean squared error (RMSE) over $M = 500$ Monte-Carlo replications as follows:

$$\text{Relative Bias}_i = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{\overline{Y}}_i^{(m)} - \overline{Y}_i^{(m)}}{\overline{Y}_i^{(m)}}\right) * 100 \tag{12}$$

$$\text{RMSE}_i = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\overline{Y}}_i^{(m)} - \overline{Y}_i^{(m)}\right)^2}, \tag{13}$$

where $\hat{\overline{Y}}_i^{(m)}$ is the estimated mean in small area i based on any of the methods mentioned above and $\overline{Y}_i^{(m)}$ denotes the true mean in small area i. Table 2 presents the results for the four scenarios and shows mean and median values of RMSE and relative bias averaged over small areas. Since the use of a log transformation is not reasonable in the *Normal* scenario, we omit these results. The same applies to the *EBP* without transformation for the *Log-Scale* and *Log-Scale Gamma* scenario.

Under the *Normal* scenario the *BHF* estimator and the *EBP* without transformation are, as expected, the best in terms of RMSE. For the methods based on a log-shift transformation, we see a small increase in terms of RMSE compared to the methods using untransformed data (*EBP* and *BHF*). Regarding the *Log-Scale* scenario, we find the best results for estimators using a fixed log transformation. The *BHF* estimator does not perform as well as the transformed (log and log-shift) *bc-agg*, *TNER2*, and *EBP* estimators in terms of relative bias and RMSE. For the *naive*, the *bc-naive-agg*, and especially for the *naive-agg* estimator, the relative bias and RMSE are the highest. The reason for this is the underestimation due to the first-order back-transformation bias (*naive*), the second-order back-transformation bias (*bc-naive-agg*), and the simultaneous first- and second-order back-transformation bias (*naive-agg*). In contrast, the proposed estimator $\hat{\overline{Y}}_i^{\text{trans, bc-agg}}$ reduces the back-transformation bias and leads to results that are similar to those of the EBP. The same picture emerges for the *Log-Scale Gamma* scenario. This demonstrates that the proposed methodology also works well when using skewed covariates and adapts to the corresponding shape of the synthetic part. In the *GB2* scenario, we see that the adaptive log-shift

Table 2: Summaries of relative bias and RMSEs over domains for different model-based scenarios

	Scenario	Normal		Log-Scale		Log-Scale Gamma		GB2	
Transformation	Estimator	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Relative bias [%]									
Gen. competitors No transformation	direct	-0.112	-0.090	-0.289	0.044	-0.228	-0.141	-0.081	-0.059
	BHF	0.174	0.144	6.871	6.457	3.986	2.480	0.189	0.205
	EBP	0.250	0.260					0.267	0.231
Log-shift	bc-agg	0.621	0.604	5.179	4.501	3.430	2.436	0.517	0.459
	bc-naive-agg	-0.951	-0.880	-7.873	-8.568	-8.146	-9.118	-2.869	-2.846
	naive	-0.654	-0.618	-38.301	-38.366	-22.313	-22.493	-4.082	-4.131
	naive-agg	-1.178	-1.751	-45.131	-45.121	-30.283	-30.420	-6.960	-6.936
	TNER2	-0.015	-0.005	0.626	0.983	-0.160	0.090	-0.337	-0.245
	EBP	0.259	0.322	4.287	3.446	2.786	1.846	0.193	0.138
Log	bc-agg			5.415	4.744	3.540	2.554	6.270	6.173
	bc-naive-agg			-7.745	-8.436	-8.134	-9.103	-0.293	-0.148
	naive			-38.364	-38.422	-22.331	-22.502	-5.070	-4.978
	naive-agg			-45.221	-45.214	-30.353	-30.485	-10.184	-10.049
	TNER2			0.825	1.179	-0.076	0.187	4.655	4.689
	EBP			4.515	3.672	2.887	1.953	5.685	5.580
RMSE									
Gen. competitors No transformation	direct	369.16	263.14	1666.97	1248.21	1788.30	1323.01	874.34	648.38
	BHF	197.69	161.48	1120.53	986.96	1364.04	1145.07	400.48	382.82
	EBP	198.84	165.03					403.43	386.07
Log-shift	bc-agg	200.04	163.14	853.08	743.12	1159.65	958.61	358.82	339.01
	bc-naive-agg	205.74	168.84	985.55	896.39	1414.44	1231.07	431.25	413.71
	naive	202.55	164.54	1948.82	1910.00	2207.02	2055.35	489.44	471.50
	naive-agg	218.18	182.08	2205.69	2175.61	2716.05	2581.19	668.34	652.34
	TNER2	281.54	243.25	936.89	819.29	1346.73	1120.87	417.25	406.16
	EBP	200.15	166.18	853.29	743.15	1161.87	964.22	358.80	342.83
Log	bc-agg			854.23	743.32	1160.38	959.12	685.44	692.48
	bc-naive-agg			981.42	890.86	1412.83	1229.57	433.94	425.20
	naive			1949.77	1910.87	2206.92	2055.27	567.36	561.69
	naive-agg			2207.99	2177.83	2719.66	2584.76	906.22	907.40
	TNER2			936.29	818.26	1346.76	1122.15	639.24	647.17
	EBP			853.56	742.47	1161.86	964.12	649.13	647.49

transformation leads to better results compared to using a fixed log transformation for all estimators. In addition, the proposed estimator (*bc-agg*) and the *EBP* under the log-shift transformation show the best results in terms of RMSE.

Overall, the simulation study leads to the following conclusions: 1) as expected, when using auxiliary information from population micro-data the *EBP* estimator performs better than the proposed estimator that is based only on aggregated population-level auxiliary information. However, despite not having access to population micro-data, the loss in efficiency is not high. 2) The proposed bias-corrected estimator reduces the relative bias compared to all three types of naive back-transformed estimators.

5.3 Performance of the bootstrap MSE estimator

We now turn our attention to the performance of the proposed parametric bootstrap presented in Section 4. We investigate the behaviour of the MSE estimator of $\hat{\overline{Y}}_i^{\text{trans, bc-agg}}$ (10) under the previous four scenarios. The proposed MSE estimator is evaluated using the estimated RMSE with $B = 500$ bootstrap replications introduced in Section 4 and the empirical RMSE (13) over $M = 500$ Monte-Carlo replica-

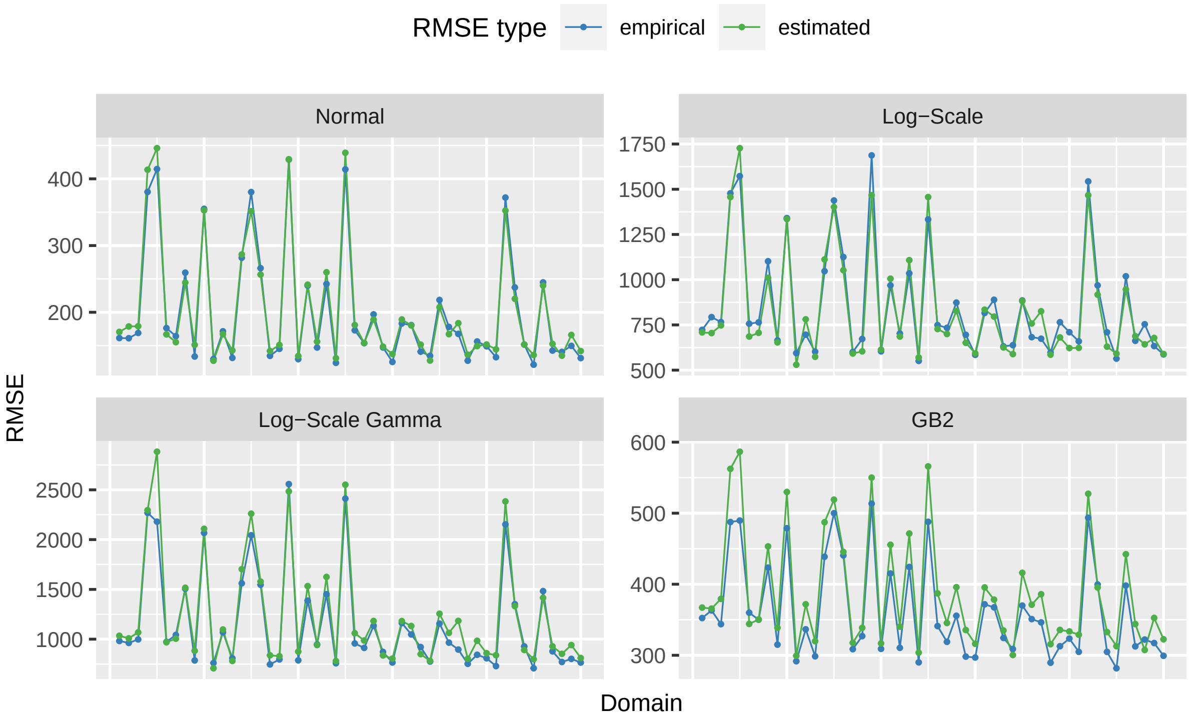


Figure 4: Estimated and empirical domain-specific RMSEs using the proposed estimator (10) with log-shift transformation in model-based simulations

tions, which we consider to be the true one. The properties of the proposed MSE estimator for each area i are assessed using the following two measures

$$\text{Relative Bias RMSE}_i = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\text{MSE}_{\text{est}_i}^{(m)}} - \text{RMSE}_{\text{emp}_i}}{\text{RMSE}_{\text{emp}_i}} * 100 \tag{14}$$

$$\text{Relative RMSE RMSE} = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\sqrt{\text{MSE}_{\text{est}_i}^{(m)}} - \text{RMSE}_{\text{emp}_i}\right)^2}}{\text{RMSE}_{\text{emp}_i}} * 100, \tag{15}$$

where $\text{MSE}_{\text{est}_i}^{(m)}$ is the estimated MSE for area i in Monte-Carlo replication m and $\text{RMSE}_{\text{emp}_i}$ is the empirical RMSE over $M = 500$ Monte-Carlo replications.

Table 3 shows the median and mean values of relative RMSE and relative bias over areas and Monte-Carlo replications of the proposed MSE estimator for the bias-corrected estimator (10) with log and log-shift transformation. For all scenarios, the proposed bootstrap MSE estimator under the log-shift transformation has reasonably low relative bias. Figure 4 shows how well the estimated RMSE tracks the empirical RMSE under log-shift transformation over the domains. These plots suggest that the estimated RMSE follows the empirical RMSE well in all four scenarios. For both *Log-Scale* scenarios, the estimated MSE under the log transformation is comparable to the corresponding MSE for the log-shift transformation.

Table 3: Performance of the MSE estimator for the proposed estimator (10) in model-based simulations

	Normal		Log-Scale		Log-Scale Gamma		GB2	
Transformation	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Relative bias RMSE [%]								
Log-shift	2.2514	2.5114	-2.1263	-2.6910	6.0185	5.3279	7.9957	8.6105
Log			-1.6072	-1.9258	6.2006	5.4005	-20.8445	-20.3301
Relative RMSE RMSE[%]								
Log-shift	6.4549	5.7627	30.3741	29.7741	15.2840	15.0982	13.4362	12.9401
Log			30.5179	29.9245	15.3057	15.2126	28.0718	25.8790

6 Design-based simulation study

In Section 5 we evaluated the performance of the point estimates and the MSE estimates in model-based simulations. In this section, we conduct a design-based simulation to assess the behaviour of different estimators in a close-to-reality environment when the true data generation mechanism is unknown. The study is based on the Mexican census of 2010. The census includes a variable closely related to income (*the earned income from work*) which is used as dependent variable in the study. Furthermore, different continuous and categorical variables are available in the census which may serve as potential auxiliary information in the models. After the description of the setup of the design-based simulation study, we discuss the performance of the proposed estimator and alternative competitors.

The initial data for the design-based simulation are population micro-data from the 2010 Mexican census from the State of Mexico. The State of Mexico consists of 125 municipalities which are the areas of interest. We draw $M = 500$ samples by stratified random sampling from the fixed census/pseudo-population following the design of the Mexican Household Income and Expenditure Survey (ENIGH) survey from 2010. The stratum is specified by the 125 municipalities. The sample size is 2748 as in the ENIGH survey with a minimum of 3 and a maximum of 527 (median: 21). The variable of interest is the earned income from work per capita (*inglabpc*) measured in Mexican pesos. As auxiliary data we use the following 6 continuous and categorical covariates:

- percentage of employees who are older than 14 years in the household;
- the highest degree of education completed by the head of household;
- the social class of the household;
- the percentage of income earners and employees in the household;
- the total number of communication assets in the household;
- the total number of goods in the household.

Modelling *inglabpc* by these covariates leads to a conditional coefficient of determination (Nakagawa and Schielzeth, 2013) for the NER model ranging from 49.6% to 59.2% for the data-driven log-shift transformation over the Monte-Carlo replications. Using the log transformation, the conditional coefficient of determination is between 46.2% and 57.6% and under no transformation the range of the conditional coefficient of determination is 26.9% to 47.1%. We evaluate the different point estimates by using the relative bias (12) and the RMSE (13) over $M = 500$ samples from the fixed population. As in the model-based simulation, the same competitors are compared with the proposed bias-corrected estimator $\hat{\bar{Y}}_i^{\text{trans, bc-agg}}$ (10) (*bc-agg*). In contrast to the model-based simulation, we do not show the re-

Table 4: Performance (in terms of the distribution of relative bias and RMSE) of estimators of mean income over municipalities in design-based simulation study

Transformation	Estimator	$Q_{0.1}$	$Q_{0.25}$	Mean	Median	$Q_{0.75}$	$Q_{0.9}$
Relative bias [%]							
Gen. competitors No transformation	BHF	-6.2471	-1.2958	7.2244	6.7426	14.6169	24.9505
	EBP	-6.7169	-1.1130	7.1482	6.3409	15.5975	24.3026
Log-Shift	bc-agg	-5.9052	-2.1816	6.9543	5.2935	12.3645	24.5303
	bc-naive-agg	-22.3306	-15.8239	-8.3422	-9.3838	-2.5686	7.3689
	naive	-21.0108	-17.4476	-9.9295	-10.9967	-4.9272	4.9207
	naive-agg	-34.9229	-29.4864	-23.2362	-23.8893	-18.6247	-10.3107
	EBP	-6.0973	-1.0638	7.3163	5.1184	13.8139	24.5727
Log	bc-agg	-3.3077	-0.7194	8.1014	6.4667	12.9045	25.3294
	bc-naive-agg	-22.3074	-15.7831	-8.6267	-9.3442	-3.3046	7.0525
	naive	-22.0931	-18.6291	-11.8209	-13.2609	-7.3448	2.3839
	naive-agg	-36.7214	-31.5616	-25.7840	-26.4056	-21.4856	-13.2655
	EBP	-3.8318	-0.0302	8.5413	6.2421	14.3214	25.7412
RMSE							
Gen. competitors No transformation	BHF	91.2780	123.0972	227.9718	193.3669	281.1682	366.0222
	EBP	82.9916	137.4112	228.5845	185.2426	296.9533	374.2862
Log-Shift	bc-agg	97.8038	128.1649	215.9485	175.2353	262.5073	372.1697
	bc-naive-agg	63.0396	119.0798	303.6137	208.9238	355.9612	595.5469
	naive	65.6649	126.2201	300.1576	224.3113	378.9602	604.9108
	naive-agg	149.1410	297.8022	533.3376	458.4745	653.1614	912.5232
	EBP	97.8839	123.8100	224.0689	180.8314	266.7279	388.3502
Log	bc-agg	109.3511	139.6032	224.3459	189.3673	268.7067	372.1409
	bc-naive-agg	64.1356	126.7677	305.2655	217.3750	350.9832	599.1836
	naive	70.1345	137.8691	324.5263	256.2682	417.7256	641.3092
	naive-agg	193.8965	349.7551	579.3393	496.0587	702.0675	960.4961
	EBP	115.0639	141.9267	232.5953	200.0447	274.2513	395.7143

sults for the *TNER2* estimator due to numerical instability associated with this estimator in the case of the design-based simulation.

Table 4 presents summary statistics of relative bias and RMSE over municipalities. The proposed estimator (*bc-agg*) leads to comparable results in terms of relative bias as the *EBP* under the same transformation, which requires population micro-data. Using naive back-transformed estimators (*naive*, *naive-agg*, and *bc-naive-agg*) lead on average to higher absolute values for the relative bias compared to the proposed bias-corrected estimator. In particular, the *naive-agg* estimator that ignores both bias-corrections performs worst. If we compare the proposed estimator with the methods using untransformed data (*EBP* and *BHF*), we observe that the proposed estimator leads on average to smaller absolute values in terms of relative bias. Regarding the RMSE, the proposed bias-corrected estimator (*bc-agg*) is almost as efficient as the *EBP* - despite the *bc-agg* estimator is using less information. All three naive estimators lead on average to high values in terms of RMSE. The general competitors build on untransformed data (*BHF* and *EBP*) does not keep up in terms of efficiency to the adaptive log-shift transformed *bc-agg* estimator. Comparing the untransformed estimators and the log transformed *bc-agg* estimator in detail, we note that in general neither estimator has superior performance over the other. This could be at least partly explained by the higher relative bias of the log transformed estimator compared to the log-shift transformed one. Working with real-world income data as is the case in the design-based simulation, we recognize that the adaptive log-shift transformation adjusts to the underlying data and therefore yields better results than the fixed log transformation (bias and uncertainty). These results are in line with the

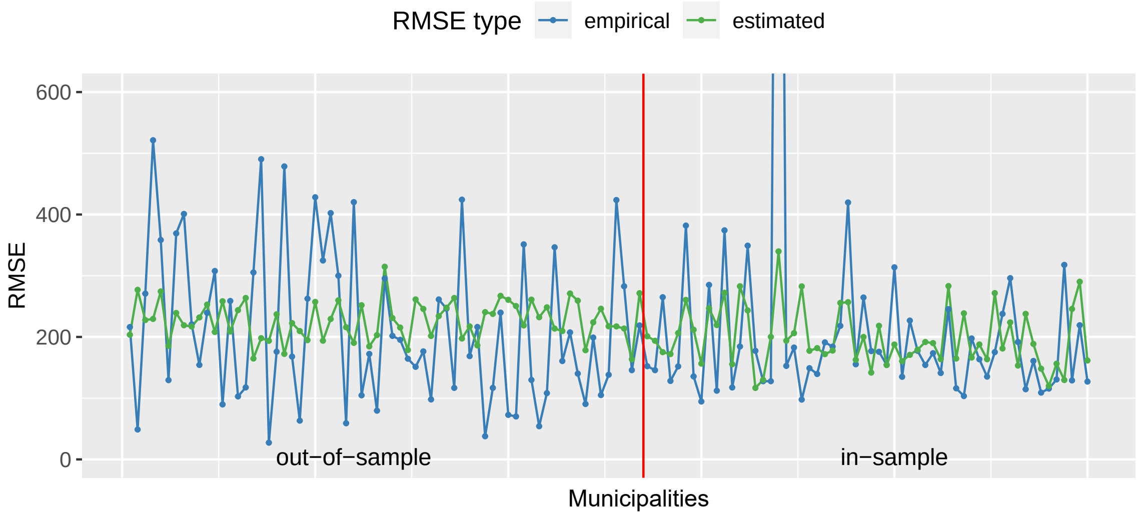


Figure 5: Estimated and empirical municipal-specific RMSE using the proposed estimator (10) with log-shift transformation in design-based simulation study

results for the *GB2*-scenario from the model-based simulation study and the findings by Rojas-Perilla *et al.* (2020) and Walter *et al.* (2021).

We now turn our attention to the behaviour of the MSE estimator - presented in Section 4 - of the proposed $\hat{\overline{Y}}_i^{\text{trans, bc-agg}}$ with an adaptive log-shift transformation. The proposed MSE estimator is evaluated in Figure 5 by using the estimated RMSE with $B = 500$ bootstrap replications and the empirical RMSE (13) over $M = 500$ samples from the fixed population. The figure indicates that the estimated RMSE tracks the empirical RMSE well especially for the in-sample municipalities. In design-based simulations, the estimation of out-of-sample municipalities is always a particular challenge. For the out-of-sample municipalities, the estimated RMSE has on average the same order of magnitude as the empirical one, but the tracking is not quite as good as for the in-sample municipalities.

In summary, the design-based study shows that the proposed point estimator performs almost on the same level as the *EBP* which requires micro population-level covariates and better than all types of naive back-transformed estimators. Furthermore, the quality of it uncertainty estimator could be demonstrated.

7 Application: Estimating income in Germany using the SOEP data

In this section we present estimates of the mean gross individual income for the 96 German RPRs using the SOEP data and aggregated auxiliary census information. The results are based on the proposed bias-corrected estimator (10) under the NER model with a log-shift transformation. MSE estimation is conducted with the parametric bootstrap we presented in Section 4 with $B = 500$ bootstrap replications. A detailed description of the survey and census datasets is given in Section 2. The use of the log-shift transformation is motivated using the previously discussed model diagnostics from Section 2.

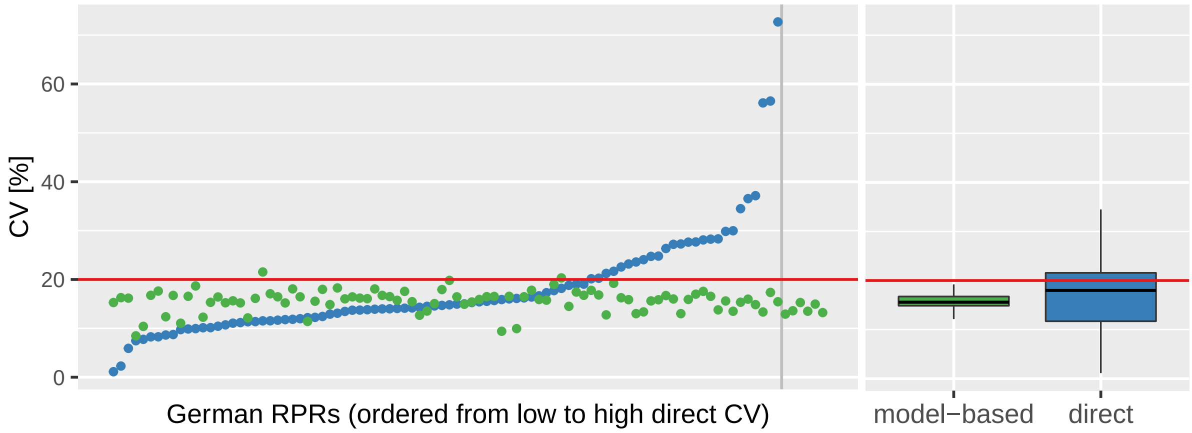


Figure 6: Area-specific CVs for the direct (blue) and the proposed model-based (green) estimates, ordered from low to high CVs for direct estimates and the associated boxplots. The grey line in the left plot separates the non-sampled areas. The red line marks the 20%-threshold for defining reliable estimates.

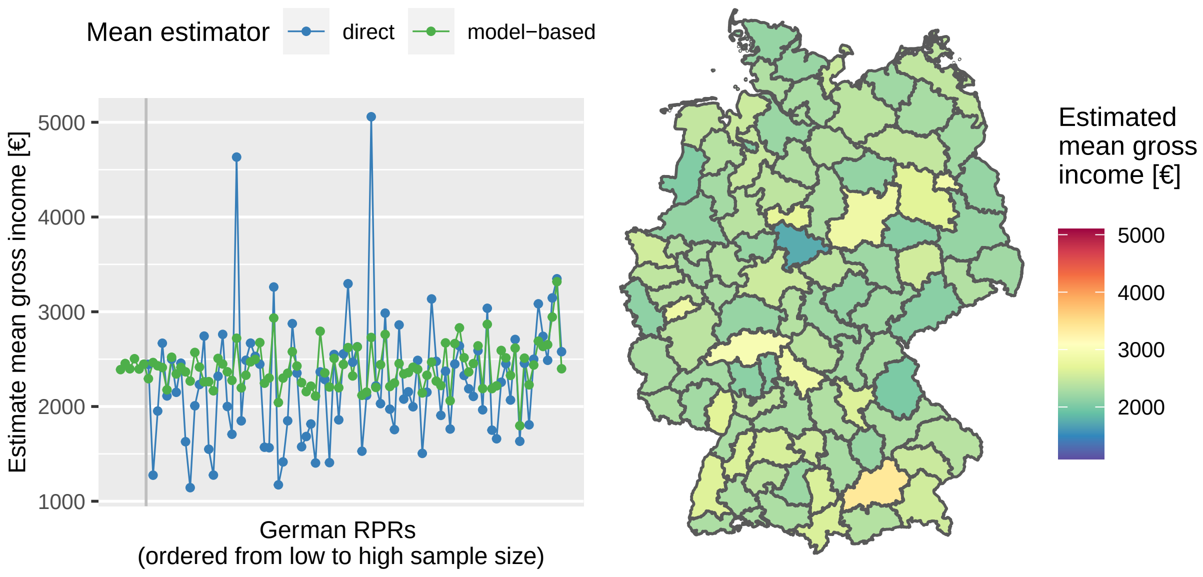
7.1 Gain in accuracy

The accuracy of the proposed estimator (10) is assessed by the estimated MSE. The variance of the direct estimates is obtained by using calibrated bootstrap variances (Alfons and Templ, 2013) which accounts for the survey design implemented in the R package emdi (Kreutzmann *et al.*, 2019). CVs for the model-based and direct estimates are computed using the point estimates and the corresponding estimates of MSE and variance estimates. As mentioned in Subsection 2.1, 26 CVs for the direct estimates exceed the 20%-threshold and 6 areas (RPRs) are out-of-sample. In comparison, only 2 CVs based on the proposed model-based estimator exceed the 20% threshold. Figure 6 shows area-specific CVs for both methods. We observe that the CVs of the model-based estimates, based on aggregated population-level covariates, are on average smaller compared to the CVs of the direct estimates (mean of the model-based CVs 15.57% vs. mean of the direct CVs 17.99%) and have a smaller interquartile range. As expected, especially for areas where the direct estimates are not reliable due to small sample sizes, the model-based estimates have improved accuracy. All in all, extreme CVs are prevented with our method and good results with higher accuracy are obtained especially for areas with small sample sizes. From these results we can conclude that the proposed approach helps with deriving improved small area estimates.

7.2 Discussion based on the application results

Figure 7 (b) shows the regional distribution of the estimated mean gross individual income with the proposed estimator (10). This map can be compared to the map of direct estimates in Figure 1 (a). It is immediately apparent that the unrealistically high range of average individual income across RPRs obtained from the direct estimates no longer exists in the model-based estimates. The line plot in Figure 7 (a) shows more clearly the relation between the proposed model-based and the direct estimates and the impact of shrinkage for areas with small sample sizes. In addition, Figure 8 and Table 6 in the appendix provide information on the mean gross individual income estimated with the *TNER2* method (Li *et al.*, 2019) for German RPRs. Note that no MSE estimator exists for *TNER2*, therefore only point estimates are shown.

The regions around economically strong cities in the West, for instance Munich and Frankfurt are the



(a) (b)

Figure 7: (a) The proposed model-based estimates (10) and direct estimates for mean gross individual income per month [€] for German RPRs ordered from low to high sample size and (b) spatial representation of corresponding model-based estimates. The grey line in the left plot separates out-of-sample areas from in-sample areas. For the former only model-based estimates are available.

RPRs with the highest mean gross individual income estimates. The other end of the income distribution includes regions in the Ruhr Area, which was affected by the breakdown of the steel industry. The lowest mean gross individual income was estimated for RPR Göttingen. Looking at the survey data, it seems that many students are in the sample (the city Göttingen within the RPR has a student rate of around 20%) and they often report low incomes under 1000€. The map shows a small difference between Eastern and Western Germany. Lower mean gross individual incomes are estimated for the eastern German states of Saxony and Thuringia. However, the RPRs in the East with the highest income report on average higher income than the West German regions with lowest income estimates. The difference between East and West Germany is considerably smaller in the model-based estimates than in the direct estimates in Figure 1 (a). Unfortunately, no other meaningful variables are available from the census that have the same definition as the variables in the SOEP data. Variables such as highest degree or work experience would be important in order to improve the predictive power of the models.

8 Conclusion

In this paper we investigate the estimation of the small area mean for income under transformations. In particular, we propose a small area estimator when only aggregate population-level auxiliary information is available. Related literature assumes access to registers or census auxiliary micro-data (Karlberg, 2000; Chandra and Chambers, 2011; Molina and Martín, 2018) which is a limitation for data analysts. From a methodological point of view, we investigate bias-correction in the case of log and log-shift transformations under aggregated population-level auxiliary information. If limited auxiliary information is present, we propose to use KDE to approximate the back-transformed totals. We don't make any parametric assumptions about the shape of the covariates. Instead non-parametric small area KDE is used to obtain an estimator for the synthetic part of the model which is used in order to reduce the second-order back-transformation bias. We further explore the use of a parametric bootstrap for estimating the MSE that captures the additional uncertainty due to the transformation and KDE. Model-based and design-based simulations are used to explore the properties of the proposed point and MSE estimators. The proposed point estimator performs comparably to the EBP under transformations (Molina and Rao, 2010; Rojas-Perilla *et al.*, 2020) that uses micro population-level auxiliary information and leads to more efficient results compared to the *TNER2* estimator of Li *et al.* (2019) that uses aggregated population-level auxiliary information. The proposed bias-corrected estimator outperforms naive back-transformed estimators and general competitors where no transformation is needed.

There are research questions that we do not investigate in the paper and are left open for further research: First, the proposed small area estimator doesn't allow for the use of survey weights in estimation, which carries risks if the assumption of non-informative sampling does not hold after conditioning on the covariates. Approaches to allow for survey weights have been proposed by Pfeffermann *et al.* (1998); Rabe-Hesketh and Skrondal (2006); Pfeffermann and Sverchkov (2007); You and Rao (2002); Guadarrama *et al.* (2018) and Burgard and Dörr (2021). Extending these approaches to allow for the use of adaptive transformations in the context of limited access to population-level auxiliary information is an open research problem. Second, in the current paper we focus on estimating small area averages. An extension of the proposed approach to estimating linear and non-linear income indicators would be valuable in obtaining a more detailed picture of the spatial distribution of income and wealth for evidence-based policymaking when population micro-data are not available. Third, we propose a parametric bootstrap MSE for quantifying the uncertainty of the small area estimates. Developing analytic MSE estimators similar to the one proposed in Molina and Martín (2018) and assessing the theoretical properties of the proposed bias-correction term offer additional avenues for future research.

Acknowledgements

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

Appendix

Supporting information for Section 4

Derivation of the area-specific distribution for each replication b in (11): Following González-Manteiga *et al.* (2008) bootstrap populations are built on a superpopulation model with $u_i^{(b)}$ and $e_{ij}^{(b)}$ following a normal distribution. Therefore, for bootstrap replication b and area i, $y_{ij}^{*(b)}$ is normally distributed with

$$\begin{aligned} E\left[y_{ij}^{*(b)}|\mathbf{y}_s^{(b)}, \mathbf{X}_s, u_i^{(b)}\right] &= \frac{1}{N_i}\sum_{j=1}^{N_i}\left[\mathbf{x}_{ij}^T\hat{\beta} + u_i^{(b)} + e_{ij}^{(b)}\right] \\ &= \overline{\mathbf{x}}_i^T\hat{\beta} + u_i^{(b)} \end{aligned}$$

and

$$
\begin{aligned}
Var\left(y_{ij}^{*(b)}|\mathbf{y}_s^{(b)},\mathbf{X}_s,u_i^{(b)}\right) &= Var\left(\mathbf{x}_{ij}^T\hat{\beta}+u_i^{(b)}+e_{ij}^{(b)}\right)\\
&= \frac{1}{N_i}\sum_{j=1}^{N_i}\left[\left(\mathbf{x}_{ij}^T\hat{\beta}+u_i^{(b)}+e_{ij}^{(b)}-\bar{\mathbf{x}}_i^T\hat{\beta}-u_i^{(b)}\right)^2\right]\\
&= \underbrace{\frac{1}{N_i}\sum_{j=1}^{N_i}\left[\left(\mathbf{x}_{ij}^T\hat{\beta}-\bar{\mathbf{x}}_i^T\hat{\beta}\right)^2\right]}_{=\sigma^2_{i,\mathbf{x}^T\hat{\beta}}}+\underbrace{\frac{1}{N_i}\sum_{j=1}^{N_i}\left[e_{ij}^{(b)2}\right]}_{\approx\hat{\sigma}_e^2}+2\underbrace{\frac{1}{N_i}\sum_{j=1}^{N_i}\left[e_{ij}^{(b)}\left(\mathbf{x}_{ij}^T-\bar{\mathbf{x}}_i^T\right)\hat{\beta}\right]}_{=0}.
\end{aligned}
$$

Additional figures and tables

Table 5: Model coefficients for the chosen linear mixed model with log-shift transformed ($\lambda = 358.73$) response variable (mean gross individual income) using the SOEP data

Fixed effects	Levels	Coefficient		Std. error
Constant		7.859	***	(0.069)
Sex	male			
	female	-0.534	***	(0.024)
Age (years)		0.006	***	(0.001)
Position in household	married			
	marriage-like	0.144	***	(0.053)
	single parent	0.051		(0.064)
	child	-0.217	***	(0.055)
	living alone	0.061	*	(0.034)
Employment status	employed (paying national insurance)			
	civil servants	0.236	***	(0.050)
	unemployed	-0.228		(0.142)
	other	-0.622	***	(0.032)
Tenant or owner	owner			
	tenant	-0.119	***	(0.027)
Migration background	non			
	direct	-0.129	***	(0.035)
	indirect	0.017		(0.044)
Random effects		Estimated variance		
RPRs ($\hat{\sigma}_u^2$)		0.0156		
Residuals ($\hat{\sigma}_e^2$)		0.2694		

Note:*p<0.1; **p<0.05; ***p<0.01

Table 6: Summaries for *direct*, model-based (with the proposed method ((10)), and *TNER2* estimates for the mean gross individual income per month [€] over German RPRs.

Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
direct	1144	1824	2244	2267	2554	5059
model-based (bc-agg)	1752	2220	2324	2373	2501	3381
TNER2	1720	2004	2152	2159	2290	2992

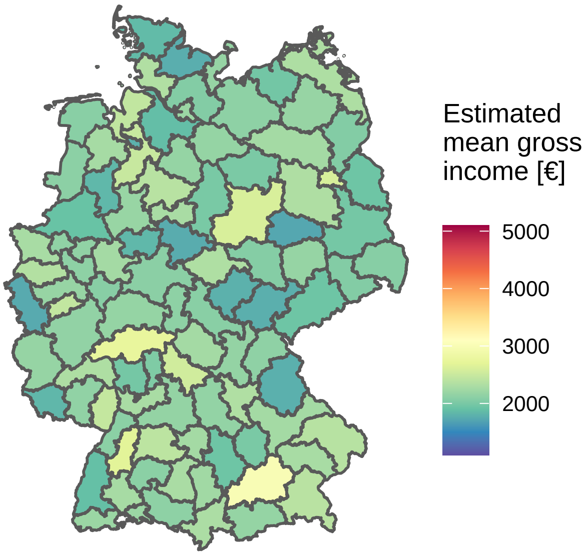


Figure 8: Spatial representation of *TNER2* estimates for mean gross individual income per month [€] for German RPRs

References

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

Battese, G. E., Harter, R. M. and 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**, 28–36.

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

Burgard, J. P. and Dörr, P. (2021) Generalized linear mixed models with crossed effects and unit-specific survey weights. *Journal of Computational and Graphical Statistics*, forthcoming.

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

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

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

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

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

Fay, R. E. and 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**, 269–277.

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

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

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

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

Görzig, B., Gornig, M. and Werwatz, A. (2008) East Germanys wage gap: a non-parametric decomposition based on establishment characteristics. *Economics of Transition*, **16**, 273–292.

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

Gurka, M. J., Edwards, L. J., Muller, K. E. and Kupper, L. L. (2006) Extending the box–cox transformation to the linear mixed model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**, 273–288.

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

Karlberg, F. (2000) Population total prediction under a lognormal superpopulation model. *Metron*, **58**, 53–80.

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

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

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

Kroh, M., Kühne, S., Siegers, R. and 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. and Zhang, R. (2019) Small area estimation under transformed nested-error regression models. *Statistical Papers*, **60**, 1397–1418.

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

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

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

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

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`).

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

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

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

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

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

R Core Team (2020) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

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

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

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

Scott, D. W. (2015) *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

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

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

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

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

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

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T. and 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**, 927–979.

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

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

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

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