The R Package **saeTrafo** for Estimating unit-level Small Area Models under Transformations

Nora Würz*

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

Abstract

The R package **saeTrafo** provides new statistical methodology for the estimation of small area means using unit-level models under transformations. The method of Würz *et al.* (2022) enables the use of unit-level models dealing with both limited auxiliary data (often the only source of data due to confidentiality agreements) and skewed distributed dependent variables like income (by using transformations such as the log or data-driven log-shift). In addition to the implementation of the new methodology, **saeTrafo** provides established methods for unit-level models under transformations, allowing further applications and comparisons. It is of advantage that the most suitable method is automatically selected and uncertainty estimates are easily offered. In addition, tools for creating plots (model validation and estimator evaluation), visualisation on maps and exporting to Excel and OpenDocument `Spreadsheets` are provided. The functionalities of the package are demonstrated with exemplary data based on Austrian income and living conditions.

Keywords: official statistics, survey statistics, small area estimation, nested error regression model, transformations

1 Introduction

For evidence-based policymaking, reliable knowledge of the spatial distribution of important variables like income is essential. As sample sizes are small at a high-resolution spatial scale of interest, direct estimates from surveys at this scale are likely to be unreliable. Small area estimation (SAE) methods are a promising and widely used approach to overcome this problem (Pfeffermann, 2013; Rao and Molina, 2015; Tzavidis *et al.*, 2018). One predominant approach - for estimating the averages in small areas - is the nested error regression (NER) model proposed by Battese *et al.* (1988) that borrows strength by using auxiliary information from a census. The starting point for this model is the availability of survey data at the individual-level. For the census data, aggregates at the spatial scale of interest are sufficient. As small area models often rely on linear mixed models, the normality assumption for the error terms has to be satisfied. However, in a variety of real-world examples, this assumption is hard to meet. Especially skewed variables, like income and consumption, can often not be adequately described by the available auxiliary variables and lead to error terms where normality assumptions are rejected. One promising approach satisfying 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. When a back-transformation to the original scale is needed, a general problem is the bias-correction. Berg and Chandra (2014) suggest an estimator with minimal mean squared error (MSE). For this estimator, Molina and Martín (2018) develop an analytical MSE estimator. It requires auxiliary information from population micro-data to correct the bias caused by the back-transformation, which is a strong limitation for data analysts. Especially in countries with high data confidentiality standards, access to individual data from the census is usually not possible. For this need, Würz *et al.* (2022) proposed methodology for estimating small area means based on the transformed NER model, if only aggregate population-level auxiliary information is available. Their approach presents an appropriate bias-correction that is necessary due to the back-transformation in the absence of population micro-data. It abstains from any parametric assumptions about the auxiliary variables and instead uses aggregate

statistics (means and covariances) and kernel density estimation (KDE) to resolve the issue of not having access to population micro-data. The authors introduce a parametric bootstrap MSE estimator that captures the uncertainty caused by the use of transformations and KDE. Alternatively, Li *et al.* (2019) propose another method relying on the smearing approach of Duan (1983) but without introducing an MSE estimator. For the second major class of small area models for estimating means - the area-level models (Fay and Herriot, 1979) - aggregated survey and population data are sufficient to determine small area means. In addition, considerable research has been done for area-level models on the application of transformations: 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 the data-driven dual power transformations. However, this model class only employs aggregates from survey data. If the user has access to individual survey data, it would be desirable to account for this finer level of survey information by applying unit-level models.

For the estimation of small area means and indicators, several software packages exist. In the following, the R software packages (R Core Team, 2022) for estimating unit-level SAE models are briefly described: the package **rsae** (Schoch, 2014) focuses on robust estimation for both unit- and area-level SAE models but do not offer transformations. Both models are also available in the R package **JoSAE** (Breidenbach, 2018) or **rhnerm** (Sugasawa, 2016). They focus on the estimation under heteroscedasticity. The R package **hbsae** (Boonstra, 2022) fits both models by maximum likelihood or hierarchical Bayesian approaches. Like the previously listed R packages, **mcmcsae** (Boonstra, 2021) also does not provide the possibility for the use of transformations. It deals with correlated random effects for both unit- and area-level models and uses markov chain monte carlo simulations. The R package **sae** (Molina and Marhuenda, 2015) offers unit-level models together with a variety of area-level models. On the one hand, it provides the classic NER model (function: `eblupBHF`). On the other hand, a NER model with transformations (box-cox and power transformation (Box and Cox, 1964)) is available, but micro-population auxiliary data is required (function: `ebBHF`). For both models, bootstrap MSEs are available. However, it is important to emphasise that `ebBHF` requires population micro-data, which is a strict limitation for data analysts. A package providing transformations for SAE methods is the **emdi** package (Kreutzmann *et al.*, 2019). It offers the area-level model and the method of Molina and Rao (2010), which requires individual census data.

The structure of **saeTrafo** is closely oriented on that of the R package **emdi** (Kreutzmann *et al.*, 2019). This means that **saeTrafo** offers similar input arguments and generic functions. The main focus of **saeTrafo** lies on making the new methodology by Würz *et al.* (2022) publicly available to enable the use of transformations (log transformation and data-driven log-shift transformation) under limited auxiliary data for unit-level small area models. The relevance is justified by data confidentiality because in developed countries like Germany, 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. Furthermore, the use of transformations is essential to meet the assumptions on the error terms. Additionally, **saeTrafo** offers further methodology in a user-friendly way: the well-known model from Battese *et al.* (1988) (without transformations), the bias-corrected estimator from Molina and Martín (2018) (which requires population micro-data), and a first-order bias-corrected estimator in the presence of aggregated population data. Depending on the used data and transformation **saeTrafo** automatically selects the appropriate method. Furthermore, the user benefits from the simple determination of the uncertainty via the main function. Some uncertainty estimates rely on bootstrap procedures. For that, **saeTrafo** supplies a parallelization option to reduce running time. Moreover, it offers well-known and SAE-specific generic functions enabling the automatic generation of plots for model diagnostics, the comparison to a direct estimator via plots, the visualization of the estimates on a map, and the easy export of the results. As the relevant graphics are generated directly within the package and personalisation options exist, it simplifies the work flow for the user.

The rest of the paper is structured as follows: Section 2 introduces the estimation methods. In Section 3, the Austrian dataset which is used to illustrate the package is described. The functionalities of **saeTrafo** are presented in Section 4. This section gives a general overview on the main function `NER_Trafo`, demonstrate this function on exemplary Austrian data, and presents generic functions for the corresponding S3 object. Section 5 outlines further potential extensions.

2 Statistical methods

The package **saeTrafo** focuses on the NER model of Battese *et al.* (1988), which uses unit-level sample data and aggregated population-level auxiliary information. For a general overview on SAE, we refer to Rao and Molina (2015) or Tzavidis *et al.* (2018). This section presents the theoretical background starting from the classical NER model to the methodology from Würz *et al.* (2022).

2.1 The nested error regression model

Throughout the paper, a 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 response y_{ij} is available for every unit in the sample s which consists of n units partitioned into sample sizes $n_1, n_2, ..., n_D$ for each area. With s_i / $\overline{s}_i$ we refer to the in-sample/out-of-sample units in area i. The vector $\mathbf{x}_{ij} = (1, x_1, x_2, ..., x_p)^T$ contains the intercept and p explanatory variables for every unit j in the sample. These vectors are combined within the matrix $\mathbf{X}_s$. The vector $\mathbf{y}_s$ contains the response of the individuals within the sample. The NER model of Battese *et al.* (1988) 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. u_i denotes the area-specific random effect and e_{ij} is the unit-level error. They are assumed to be independent and σ_u^2 and σ_e^2 denote their variances. An out-of-sample unit is estimated as best linear unbiased prediction by $\hat{\mu}_{ij} = \mathbf{x}_{ij}^T\hat{\beta} + \hat{u}_i = \mathbf{x}_{ij}^T\hat{\beta} + \hat{\gamma}_i \left(\sum_{j \in s_i} \left(y_{ij} - \mathbf{x}_{ij}^T\hat{\beta}\right)\right)$, where $\hat{\gamma}_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2/n_i}$ denotes the estimated shrinkage factor. The target parameter is the population mean for each area i and it is estimated as the empirical best linear unbiased predictor (EBLUP) for the population area mean ($\overline{y}_i$) by

$$\begin{aligned} \hat{\overline{Y}}_i^{\text{BHF}} &= \frac{1}{N_i}\left(\sum\nolimits_{j \in s_i} y_{ij} + \sum\nolimits_{j \in \overline{s}_i} \hat{\mu}_{ij}\right) \\ &= \hat{\gamma}_i \left(\frac{1}{n_i}\sum_{j \in s_i} y_{ij} + \left(\overline{\mathbf{x}}_i - \frac{1}{n_i}\sum_{j \in s_i} \mathbf{x}_{ij}\right)^T \hat{\beta}\right) + (1 - \hat{\gamma}_i)\,\overline{\mathbf{x}}_i^T\hat{\beta}. \end{aligned} \tag{2}$$

The vector $\overline{\mathbf{x}}_i^T = \frac{1}{N_i}\sum_{j \in U_i} \mathbf{x}_{ij}^T$ contains means for the p covariates within i. **saeTrafo** uses the restricted maximum likelihood (REML) theory to estimate fixed effects and the variance components. As in the package **emdi** (Kreutzmann *et al.*, 2019), it is implemented based on the `lme` function of the package **nlme** (Pinheiro *et al.*, 2022). Note that the estimator of Battese *et al.* (1988) ($\hat{\overline{Y}}_i^{\text{BHF}}$, (2)) requires only population-level aggregates and a unit-level survey.

To estimate the uncertainty of $\hat{\overline{Y}}_i^{\text{BHF}}$ (2), Prasad and Rao (1990) propose an analytical MSE which **saeTrafo** supplies. A second possibility for determining the uncertainty are bootstrap methods offered by R packages such as **sae** (Molina and Marhuenda, 2015).

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

One-to-one transformations of the response $h(y_{ij}) = y_{ij}^*$ are a common tool to prevent violations of the model assumptions. For skewed variables, like income, this problem is typical. In order to adapt better to the data, data-driven transformations are promising for SAE (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. In **saeTrafo**, the transformation parameter λ is estimated from the sample data using the REML method as Rojas-Perilla *et al.* (2020) proposed.

Using a transformation on the response results in a model on the transformed scale:

$$h(y_{ij}) = y_{ij}^* = \mathbf{x}_{ij}^T\beta + u_i + e_{ij}, \qquad u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2) \text{ and } e_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2). \tag{3}$$

The BLUP on the transformed scale for out-of-sample units is $\mu_{ij}^* = \mathbf{x}_{ij}^T\beta + u_i$. However, in SAE applications there is interest in prediction, so the aim is to estimate the mean on the original scale.

Due to Jensen's inequality, the naive back-transformation of real convex or concave functions $h()$ don't lead to the same result as the best prediction on the original scale (Jensen *et al.*, 1906):

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

For the log and log-shift transformation, the back-transformation $h^{-1}() = \exp()$ or $h^{-1}() = \exp() - \lambda$ is convex and hence $\mu_{ij}^{\text{trans, naive}}$ underestimates $E[h^{-1}(y_{ij}^*)|\mathbf{y}_s, \mathbf{X}_s]$. In order to get bias-corrected estimates, the best prediction on the original scale is needed.

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 (3),

$$\begin{aligned}\mu_{ij}^{\text{trans, bc}} = E[h^{-1}(y_{ij}^*)|\mathbf{y}_s,\mathbf{X}_s] &= \int_{-\infty}^{+\infty} h^{-1}(x) f_{y_{ij}^*|\mathbf{y}_s,\mathbf{X}_s}(x)dx \\ &\overset{h^{-1}()=\exp()}{=} \exp\left(\mu_{ij}^* + \underbrace{\frac{\sigma_u^2(1-\gamma_i)+\sigma_e^2}{2}}_{=\alpha_i \text{ (bias-correction)}}\right).\end{aligned}$$

To the BLUP on the transformed scale (μ_{ij}^*) a bias-correction (α_i) is added before applying the back-transformation. $\mu_{ij}^{\text{trans, bc}}$ can be used to determine the bias-corrected estimator of the small area mean:

$$\hat{\overline{Y}}_i^{\text{trans, bc}} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} \hat{\mu}_{ij}^{\text{trans, bc}}\right) = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} \exp\left(\mathbf{x}_{ij}^T\hat{\beta} + \hat{u}_i + \hat{\alpha}_i\right)\right). \quad (4)$$

Molina and Martín (2018) propose for the MSE of $\hat{\overline{Y}}_i^{\text{trans, bc}}$ (4) both an analytical and a parametric bootstrap estimator. The package **saeTrafo** provides (4) and its bootstrap MSE estimator.

For $\hat{\overline{Y}}_i^{\text{trans, bc}}$ (4), out-of-sample population micro-data are needed which often causes problems with data confidentiality. Again, due to the Jensen's inequality a (second-order) bias is introduced if we use a naive back-transformation of the synthetic part (i.e., $\exp\left(\overline{\mathbf{x}}_i^T\hat{\beta}\right)$ 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

$$\hat{\overline{Y}}_i^{\text{trans, bc-naive-agg}} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \overline{s}_i} \exp\left(\overline{\mathbf{x}}_i^T\hat{\beta} + \hat{u}_i + \hat{\alpha}_i\right)\right). \quad (5)$$

Due to the use of aggregated auxiliary data, this estimator has a second-order bias. To the best of my knowledge, no MSE estimator exists for $\hat{\overline{Y}}_i^{\text{trans, bc-naive-agg}}$ (5).

The next subsection presents small area means under the transformed NER model if only aggregated population-level auxiliary information is available. Therefore, it addresses the problem of limited data access and simultaneous transformation.

2.3 Small area means under limited auxiliary information

As emphasized in the previous subsection, the estimator $\hat{\overline{Y}}_i^{\text{trans, bc}}$ (4) requires population-level auxiliary data, which often leads to confidentiality constraints. In $\hat{\overline{Y}}_i^{\text{trans, bc-naive-agg}}$ (5), a second order bias remains because aggregated auxiliary data is used instead of individual data. In contrast to this, the method of

Würz *et al.* (2022) aims to reduce 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 while using log or log-shift transformation. This method approximates $\mathbf{x}_{ij}^T\hat{\beta}$ in the absence of population micro-data to reduce the second-order bias and combines this with the first-order bias-correction (α_i) for small area means.

Kernel density estimation for the synthetic part Due to limited auxiliary information, it is not possible to obtain $\left(\sum_{j\in\overline{s}_i}\exp\left(\mathbf{x}_{ij}^T\hat{\beta}\right)\right)$ necessary for computing $\hat{\overline{Y}}_i^{\text{trans, bc}}$ (4). Würz *et al.* (2022) propose an estimation method for the unknown synthetic part ($\mathbf{x}_{ij}^T\hat{\beta}$) under limited auxiliary information. They employ a KDE approach to estimate the distribution of $\mathbf{x}_{ij}^T\hat{\beta}$. This approach has two main advantages: firstly, the method of Würz *et al.* (2022) uses univariate KDE for the synthetic part $\left(\mathbf{x}_{ij}^T\hat{\beta}\right)$ instead of multivariate KDE to estimate the joint multivariate distribution of the auxiliary variables. Since current implementations of multivariate KDEs in R are restricted to a maximum number of auxiliary variables (cf. the widely used package **ks** (Duong, 2022) only allows for up to 6 covariates), many applications especially those with categorical data very quickly reach this limit. In contrasts, univariate KDE for the synthetic part avoids this restriction. Simulation studies in Würz *et al.* (2022) show that the estimation of the synthetic part is sufficient to reduce the second-order bias. Secondly, this method does not impose any parametric assumptions on the covariates and only require aggregated population-level auxiliary information.

KDE was first mentioned by Rosenblatt (1956) and Parzen (1962). Formally, KDE estimates the density f of a sample $X = \{X_1, ..., X_n\}$ by

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right), \tag{6}$$

where the function $k()$ is the kernel and h is the bandwidth. For more details on KDE, see for example Scott (2015). **saeTrafo** employs the Epanechnikov kernel (Epanechnikov, 1969), which is implemented using the `density` function of the **stats** package. Moreover, **saeTrafo** uses the method from Sheather and Jones (1991) for bandwidth selection.

As a first step, **saeTrafo** standardizes the predictions of the synthetic part from the NER model. For area i and individual j, the standardized predicted values z_{ij} are computed by

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

This formula employs the mean and the standard deviation from the sample data predictions of the synthetic part.

Second, the package adjusts the predictions with the help of aggregated population-level auxiliary data. It uses 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. This step incorporates the aggregated information from the census, which adds the SAE component to this method. Typically, in small area applications, sample sizes differ between areas. The package distinguishes between large sample sizes - standardized data (z_{ij}) from the respective area i (conditional) is used - and small sample sizes - standardized data (z_{ij}) from all areas (unconditional) is employed. In order to distinguish between large and small sample sizes, a threshold t is defined: for small sample sizes, i.e. below the threshold ($n_i < t$) - or even for an out-of-sample area - we use the standardized data from all areas to generate adjusted data for area i. The input values for the KDE (r_{im}) arise from the standardized 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). With

$$r_{im} = z_m\,\sigma_{i,\mathbf{X}^T\hat{\beta}} + \overline{\mathbf{x}}_i^T\hat{\beta} \qquad \text{for} \begin{cases} m \in s & n_i < t \\ m \in s_i & n_i \geq t \end{cases} \tag{7}$$

we estimate the respective density using the KDE (6) for each area i. $\hat{f}_{h,i}$ denotes the resulting density for area i.

Small area means under limited auxiliary information In order to account for both types of biases 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:

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

$\hat{\mu}_{ij} = \mathbf{x}_{ij}^T\hat{\beta} + \hat{u}_i$ is defined as in the NER model. As shown above, under limited auxiliary information, the problem is reduced to determining the unknown back-transformed total (T_i). Würz *et al.* (2022) use numerical integration and the estimated density of the synthetic part $\hat{f}_{h,i}$ to determine the total $\hat{T}_i = \sum_{j=1}^{N_i} \exp\left(\mathbf{x}_{ij}^T\hat{\beta}\right) = N_i E[\exp(\mathbf{x}_{ij}^T\hat{\beta})] = N_i \int_{-\infty}^{+\infty} \exp(x)\hat{f}_{h,i}(x)dx$ from sample data and population-level auxiliary information - without using population micro-data. To achieve this, **saeTrafo** uses the package **sfsmisc** (Maechler *et al.*, 2021). The requested small area estimator of the mean is obtained by inserting the estimated back-transformed area-specific totals $\hat{T}_i$:

$$\hat{\overline{Y}}_i^{\text{trans, bc-agg}} = \frac{1}{N_i}\hat{T}_i \exp\left(\hat{u}_i + \hat{\alpha}_i\right). \tag{8}$$

For the log-shift transformation, the characteristic shift-parameter $\hat{\lambda}$ is added

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

The R package **saeTrafo** is the first package providing these estimators to the users.

Uncertainty estimation For the estimator $\hat{\overline{Y}}_i^{\text{trans, bc-agg}}$ (8) under limited auxiliary data, Würz *et al.* (2022) 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 following enumeration outlines the bootstrap procedure employed in **saeTrafo** for the log and log-shift transformation (these transformations are denoted with h).

1. Transform the data: $y_{ij}^* = h(y_{ij})$
2. Estimate $\hat{\beta}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ from sample data using model (3). In the case of the log-shift transformation, estimate $\hat{\lambda}$ 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) Build bootstrap samples on the transformed scale
   $$y_{ij}^{*(b)} = \mathbf{x}_{ij}^T\hat{\beta} + u_i^{(b)} + e_{ij}^{(b)}, \qquad \text{with } j \in s_i$$
   for all areas i and therefore determine the estimator $\hat{\overline{Y}}_i^{\text{trans, bc-agg, (b)}}$ (8) for all areas within each bootstrap replication b. Note, that λ is re-estimated within every replication b in case of the log-shift transformation.
   (c) Determine the true mean for each area i in each bootstrap replication b. Due to the lack of population micro-data for $\mathbf{x}$, an approximation of the true bootstrap mean is needed. From the available aggregated population-level values, Würz *et al.* (2022) construct an area-specific distribution on the transformed scale for each bootstrap replication b:
   $$y_{ij}^{*(b)} | \mathbf{y}_s^{(b)}, \mathbf{X}_s, u_i^{(b)} \sim \mathcal{N}\left(\overline{\mathbf{x}}_i^T\hat{\beta} + u_i^{(b)}, \sigma^2_{i,\mathbf{X}^T\hat{\beta}} + \hat{\sigma}_e^2\right), \tag{9}$$

determine $\sigma_{i,\mathbf{X}^T\hat{\beta}} = \sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p}\hat{\beta}_k\hat{\beta}_l \text{Cov}[\mathbf{x}_{ik},\mathbf{x}_{il}]}$ from known covariances and estimated regression coefficients, and take $\hat{\sigma}_e^2$ from step 2. To get the true mean ($\overline{Y}_i^{(b)}$) on the original scale, Würz *et al.* (2022) combine the distributional assumptions on the transformed scale (9) with the properties of the exponential back-transformation function $h^{-1}() = exp()$, respectively $h^{-1}() = exp() - \lambda$:

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

For data-driven log-shift transformation, the analogue 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:

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

saeTrafo offers this parametric bootstrap procedure. To increase user-friendliness, it is possible to run this MSE estimation procedure on several cores. The expected execution times are displayed to the users.

The next section describes the Austrian data while Section 4 presents the core function `NER_Trafo`. The function provides the theory from this section in a user-friendly way, and demonstrates it based on the Austrian data.

3 Data sets for illustration

The main idea of SAE is to combine survey and population (census or administrative) data to increase the accuracy of the estimated indicator of interest. Since the target variable is only provided in the survey data, additional information from the population is used to support the prediction of the target variable using linear mixed models (Rao and Molina, 2015; Tzavidis *et al.*, 2018). The package **saeTrafo** contains sample and population data to provide the users with exemplary data. The sample (`eusilcA_smp`) and population data (`eusilcA_pop`) are obtained from the package **emdi** (Kreutzmann *et al.*, 2019). The authors provide an extensive description of the data generating processed of the `eusilcP` dataset coming from the package **simFrame** (Alfons *et al.*, 2010). This household-level data set consists of synthetic Austrian European Union Statistics on Income and Living Conditions (EU-SILC) from 2006. For the package **emdi**, a spatially finer regional disaggregation was generated manually using a random assignment taking into account the different regional income-levels in Austria. The lowest regional level in this synthetic data set are the 94 Austrian districts. This population data comprises 25000 households, while there were more than 3.5 million households in Austria in 2006. The sample data is constructed by stratified random sampling and consists of 1945 households. The sample data includes 70 districts, leaving 24 areas out-of-sample. The equivalized household income (`eqIncome`) is the target variable and is only available within the sample. This variable is defined as the ratio of the total household disposable income and the equivalized household size. It was determined by the Organisation for Economic Cooperation and Development (OECD) (Hagenaars *et al.*, 1994). In the following examples, 14 covariates serve as auxiliary data: `gender`, `eqsize`, `cash`, `self_empl`, `unempl_ben`, `age_ben`, `surv_ben`, `sick_ben`, `dis_ben`, `rent`, `fam_allow`, `house_allow`, `cap_inv`, and `tax_adj`. For detailed

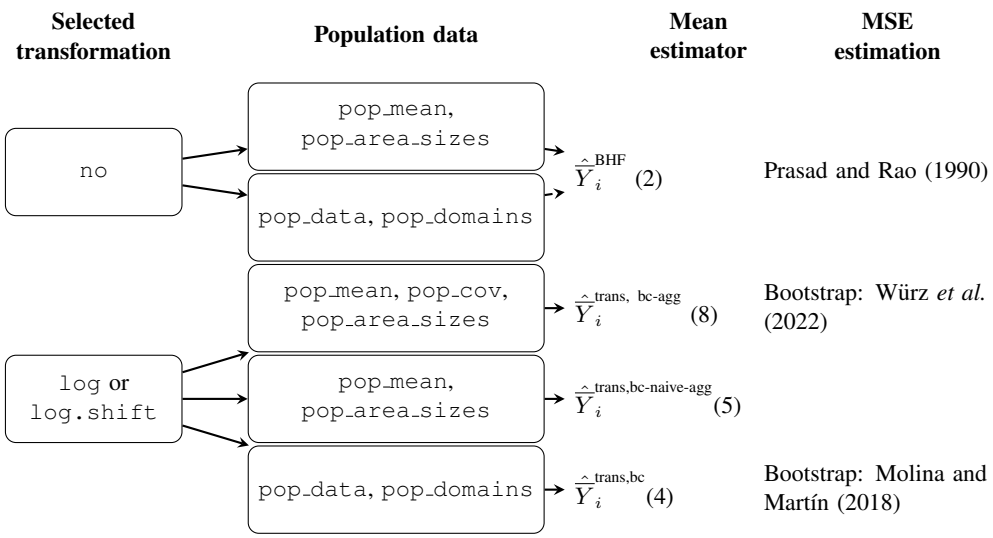

Figure 1: Overview of different estimation methods provided in function `NER_Trafo`. These estimation methods are chosen depending on the selected transformation and the type of provided population data.

information, please refer to Kreutzmann *et al.* (2019). All 14 covariates are included within the sample and the full and aggregated population data. Furthermore, the variable `district` is available in the data and represents the spatial target level.

The core function `NER_Trafo` of the package **saeTrafo** deals with different population data inputs. Figure 1 visualizes, which functions from the theory part (Section 2) applies under which population data input. To provide aggregates in a directly and user-friendly manner, `pop_area_size`, `pop_mean`, and `pop_cov` are available as data sources in the package. All three data objects are calculated from `eusilcA_pop`. Their direct availability makes it more convenient for the user to try out all functionalities of **saeTrafo**.

4 Core functionalities

This section is structured accordingly: Section 4.1 gives an overview of the main function `NER_Trafo`, Section 4.2 shows how `NER_Trafo` is applied using the exemplary data, and Section 4.3 demonstrates the possibilities of **saeTrafo**'s generic functions to analyse, visualize, and export the corresponding S3 object.

4.1 Overview NER_Trafo

The `NER_Trafo` function provides the methodology from Section 2. `NER_Trafo` has 16 input arguments, takes the different data input possibilities into account, and allows for a variety of specifications (cf. Table 1). As a minimum input, the sample data (`smp_data` and `smp_domains`), the formula object (`fixed`), and population data - either the aggregated data (`pop_area_size`, `pop_mean`, and optional `pop_cov`) or the individual data (`pop_data` and `pop_domains`) - must be entered. As **saeTrafo** uses the S3 object system, `NER_Trafo` returns an object of class `saeTrafo` and `NER` (Chambers and Hastie, 1992). The reason for assigning two classes to the object is ability to integrate further SAE models in future releases. The output object consists of ten components. In this way, the user can directly access the point estimates (`ind`), the uncertainty estimates (`MSE`), transformation parameters (`transform_param`), information on the underlying linear mixed-effects model as in the package **nlme** (Pinheiro *et al.*, 2022) (`model`), a list describing the data input (`framework`), the selected transformation (`transformation`), the method for transformation parameter estimation (`method`), the formula object (`fixed`), the function call (`call`), and number of successful bootstraps for bootstrap MSE estimation procedures (`successful_bootstraps`).

Figure 1 illustrates which estimation methods for point and MSE estimation are used under different

Table 1: Input arguments of function NER_Trafo.

Arguments	Short description	Default
fixed	Formula object with fixed effects and response variable of the NER model	
pop_area_size	Population sizes per domain	NULL
pop_mean	Population means for all fixed effects per domain	NULL
pop_cov	Population covariance matrices between all fixed effects per domain	NULL
pop_data	Census or administrative data containing all fixed effects	NULL
pop_domains	Domain identifier for population data	NULL
smp_data	Survey data comprising the fixed effects and the response variable	
smp_domains	Domain identifier for sample data	
threshold	Threshold for using pooled domain data	30
B	Number of bootstrap replications for bootstrap MSE estimation	50
transformation	Type of transformation: no, log, log-shift	log-shift
interval	Interval for estimating the optimal parameter of log-shift transformation	range of response
MSE	MSE estimation	FALSE
parallel_mode	Mode of parallelization for bootstrap MSE procedure	automatic
cpus	Kernels for parallelization for bootstrap MSE procedure	1
seed	Seed for random number generator within bootstrap MSE procedure	123

combinations of selected transformation and type of population data. If no transformation is selected, **saeTrafo** employs the classical model by Battese *et al.* (1988). Since no individual data are necessary, potentially used population micro-data are processed into aggregates in a first step. Under the log or log-shift transformation **saeTrafo** automatically selects between different methods depending on the data. **saeTrafo** uses the estimator of Würz *et al.* (2022) if population aggregates (means, covariances, and populations area sizes) are supplied in the presence of transformations. If only means and area sizes under log or log-shift transformation are present, the NER_Trafo function employs the estimator $\hat{\overline{Y}}_i^{\text{trans,bc-naive-agg}}$ (5) for which no MSE estimator exists. This estimator only corrects the first-order bias and neglects the second bias due to limited data. An alternative method - not implemented in R yet - is the estimator from Li *et al.* (2019), for which no MSE estimator exists too. If the log or log-shift transformation occur with individual population data, **saeTrafo** uses the estimator $\hat{\overline{Y}}_i^{\text{trans,bc}}$ (4) together with its bootstrap MSE. Please note, that in the cases of individual population data other packages like **emdi** (Kreutzmann *et al.*, 2019) provide further functionalities: the estimation of quantiles, inequality indicators, and further transformations (box-cox transformation (Box and Cox, 1964) and dual transformation (Yang, 2006)). These options become available in the ebp function of **emdi** which applies the method of Molina and Rao (2010). Since the ebp function is based on Monte Carlo replications, the run time is longer than for NER_Trafo.

4.2 Estimation of (transformed) nested error regression models

Synthetic Austrian EUSILC data (cf. Section 3) is used to illustrate the functionalities of **saeTrafo** and the estimation with NER_Trafo. The example demonstrates the estimation of the small area means for the equivalized household income (eqIncome) at the disaggregation level of 94 Austrian districts. The sample, population, and aggregated data are available in **saeTrafo**:

```
R> library(saeTrafo)
R> data("eusilcA_pop")
R> data("eusilcA_smp")
```

```
R> data("pop_area_size")
R> data("pop_mean")
R> data("pop_cov")
```

The data allow for easy testing of the different methods implemented and bundled in `NER_Trafo`. For illustration purposes, the example focuses on estimating $\hat{\overline{Y}}_i^{\text{trans, bc-agg}}$ (8), therefore it is sufficient to insert only aggregated population data. In addition to the point estimates, MSE estimates are calculated too, so `MSE` is set to `TRUE`. Furthermore, the setting for the `threshold` for pooled estimation (cf. (7)) is set to 50. To prevent long run times for MSE estimation the default of the number of bootstrap replications is only 50, whereby parallelization is available in the function. To obtain a more precise MSE estimate, `B` is increased to 250 in the example. The `seed` is set to 2022 to ensure reproducibility of the results.

```
R> formula <- eqIncome ~ gender + eqsize + cash + self_empl +
+    unempl_ben + age_ben + surv_ben + sick_ben + dis_ben +
+    rent + fam_allow + house_allow + cap_inv + tax_adj

R> NER_model <- NER_Trafo(fixed = formula,
+    pop_area_size = pop_area_size, pop_mean = pop_mean,
+    pop_cov = pop_cov, smp_data = eusilcA_smp,
+    smp_domains = "district", B = 250, threshold = 50,
+    MSE = TRUE, seed = 2022)
```

The R object `NER_model` is of two classes `saeTrafo` and `NER`. For this S3 object several generic functions are provided within **saeTrafo** and presented in the following section.

4.3 Generic functions

The most important generic functions of the R package **saeTrafo** (summary output, diagnostic plots, visualisation of estimates, and their export) are shown in detail. All other functionalities are only briefly introduced.

Summary of a saeTrafo object By applying the `summary` function on an object of class `saeTrafo`, R-user receive basic information and first diagnostic results. In addition to the call, small area specific characteristics (number of out-of-sample and in-sample domains, information on sample sizes, and their distribution among domains) are displayed. To assess the proportion of variance explained by the model, **saeTrafo** provides both a marginal and conditional R^2 following Nakagawa and Schielzeth (2013). The R^2s are implemented as in the **emdi**-package (Kreutzmann *et al.*, 2019) and use the **MuMIn**-package from Barton (2018). Moreover, the output shows information on the residual diagnostics for the unit-level errors (e_{ij}) and the domain-specific random effects (u_i). If a transformation is selected, **saeTrafo** calculates these diagnostics on the transformed scale and hence help to judge, if the transformation assists to meet the normality assumption of both components. The ICC relates the variances (σ_u^2 and σ_e^2) to each other. Finally, the `summary` function outputs information on the transformation and the selected parameter λ.

```
R> summary(NER_model)

Nested Error Regression Model

Call:
 NER_Trafo(fixed = eqIncome ~ gender + eqsize + cash +
    self_empl + unempl_ben + age_ben + surv_ben + sick_ben +
    dis_ben + rent + fam_allow + house_allow + cap_inv +
    tax_adj,
    pop_area_size = pop_area_size, pop_mean = pop_mean,
```

```
    pop_cov = pop_cov, smp_data = eusilcA_smp,
    smp_domains = "district", threshold = 50, B = 250,
    MSE = TRUE, seed = 2022)

Out-of-sample domains:  24
In-sample domains:  70

Sample sizes:
Units in sample:  1945
Units in population:  25000

                   Min. 1st Qu. Median      Mean 3rd Qu. Max.
Sample_domains       14    17.0   22.5  27.78571   29.00  200
Population_domains    5   126.5  181.5 265.95745  265.75 5857

Explanatory measures:
 Marginal_R2 Conditional_R2
   0.6233538      0.7054886

Residual diagnostics:
               Skewness Kurtosis Shapiro_W    Shapiro_p
Error         0.6222910 7.607189 0.9706711 1.705890e-19
Random_effect 0.4788713 2.726898 0.9737695 1.487627e-01

ICC:  0.2180689

Transformation:
 Transformation Method Optimal_lambda
      log.shift   reml       27907.57
```

The output of the example shows that the synthetic Austrian data consists of 24 out-of-sample domains and 70 in-sample domains. As the sample sizes over domains are considerably small (Median: 22.5) this is a classical small area problem. Both the marginal and conditional coefficients of determination are high with values above 62%. The normality assumption for the random effects is not rejected at a significance level of 5%. For the individual errors, this assumption is rejected with p = 1.705890e-19. The random effects contribute to around 21% of the model variance. The chosen transformation is the log-shift transformation with REML-estimated transformation parameter of $\lambda = 27907.57$.

Diagnostic plots for the nested error regression model The `plot` function provides five plots bundling the most important diagnostic information: Q-Q plots to judge the normality assumption on the error terms (cf. Figure 2a), the deviation of both the density from the normal distribution for the individual errors (cf. Figure 2b) and the random effects (cf. Figure 2c), the Cook's distance to identify outliers (cf. Figure 2d) as well as information on the optimal transformation parameter λ for the log-shift transformation (cf. Figure 2e). The `plot` function allows customized settings: the input arguments `label`, `color`, `cooks`, and `range` enable direct changes to the plots. In addition, with `gg_theme` there is the possibility of further personalisation of the plots by using the **ggplot2** package (Wickham, 2016).

```
R> plot(NER_model)
```

In the Austrian income example, the Q-Q plot (cf. Figure 2a) and the density plot (cf. Figure 2c) confirm the normality assumptions of the underlying model for the random effects. However, for the individual error term, the Q-Q plot (cf. Figure 2a) shows several outliers. The Cooks distance plot highlights three individuals as possible outliers. The last plot (cf. Figure 2e) shows the log-likelihood reaching its maximum at $\lambda = 27907.57$. This plot is only supplied for the log-shift transformation.

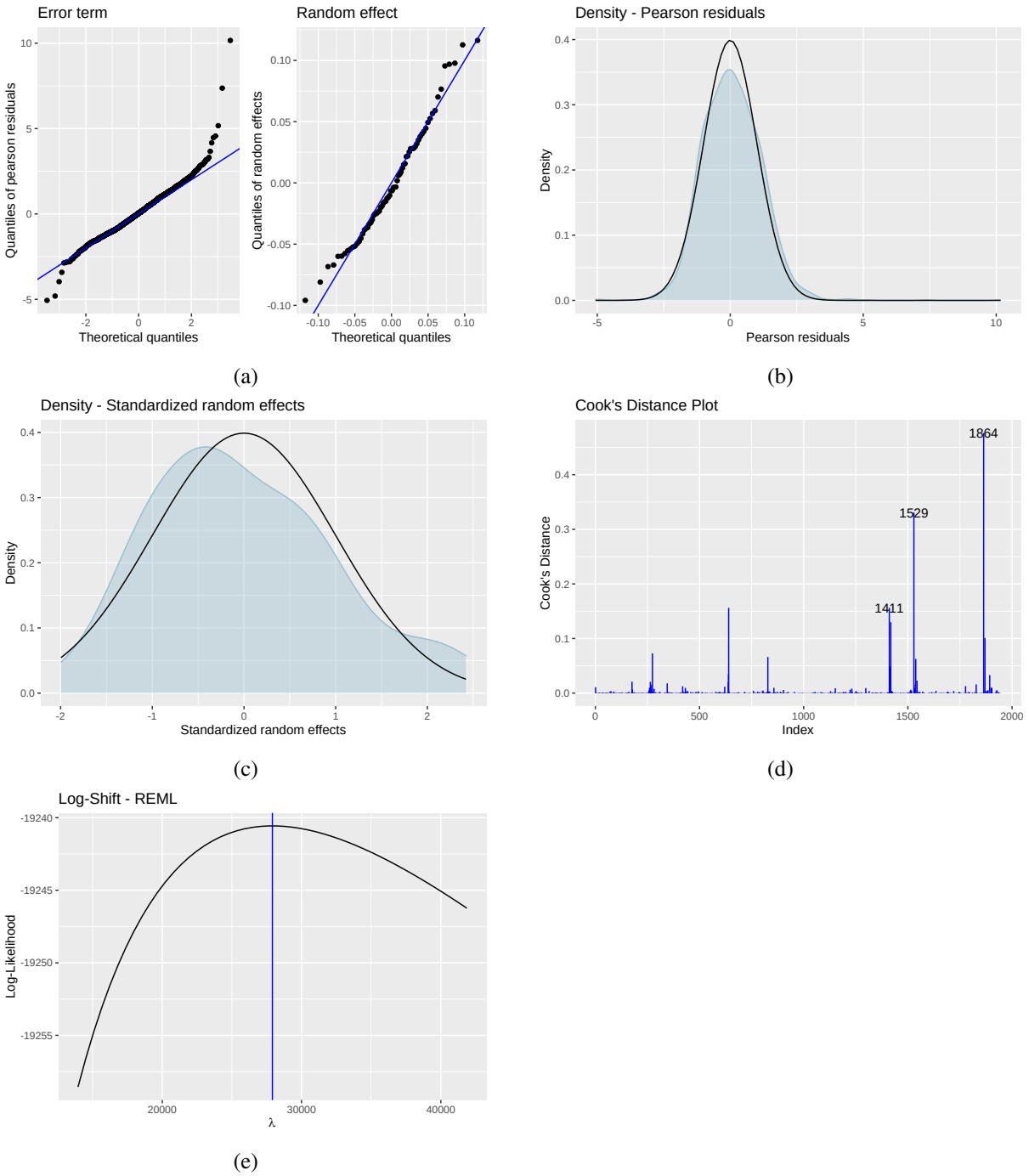

Figure 2: Diagnostic plots from generic function `plot`: Q-Q plots (a) and two density plots ((b) and (c)) to check the normality assumption for both error terms, Cook's distance plot for detecting potential outliers (d), and log-likelihood for the optimal shift parameter λ (e).

Comparing point and optional MSE/CV estimates The generic function `compare_plot` is very important for users to evaluate the quality of their model-based estimates. In SAE applications the comparison of the particular model-based estimator to the respective direct estimator is of central importance. Since **saeTrafo** does not provide a function for determining direct estimators, other packages must be utilized. Among others the **survey** package (Lumley, 2004), the **laeken** package (Alfons and Templ, 2013), and the **emdi** package (Kreutzmann *et al.*, 2019) enable the estimation of disaggregated direct estimators and their variances from a survey. Up to now, the generic function `compare_plot` works only with direct estimators from the package **emdi**. The procedure for this is shown in the exemplary code. For the comparison of point estimates, `compare_plot` returns two types of plots: a scatter plot following Brown *et al.* (2001) and a lineplot with direct and model-based domain-wise estimates. To compare the uncertainty - if `MSE` or `CV` is set to `TRUE` - `compare_plot` returns a boxplot and a scatterplot. In addition to a direct adjustment of the visualisation with `label`, `color`, `shape`, and `line_type` the argument `gg_theme` offers the possibility for further visualisation options using the **ggplot2** package (Wickham, 2016).

```
R> require(emdi)
R> library(emdi)
R> emdi_direct <- direct(y = "eqIncome",
+    smp_data = eusilcA_smp, smp_domains = "district",
+    weights = "weight", var = TRUE,
+    na.rm = TRUE)
R> detach("package:emdi", unload = TRUE)

R> compare_plot(model = NER_model, direct = emdi_direct,
+    CV = TRUE)
```

Both plots comparing direct and model-based point estimates show that the direct and model-based estimates are close to each other, as the regression line and the identity line are close to each other (cf. Figure 3a) and the model-based estimates track the direct ones (cf. Figure 3b). Furthermore, the CV is assessed in Figure 3c and 3d. As the boxplots show, the uncertainty - measured by the CV - is reduced clearly. The scatterplot which orders the domains by their sample size (from low to high) supports this impression.

Visualization of regional disaggregated estimates on a map The spatial visualisation on a map is simplified considerably by the `map_plot` function which generates maps automatically if a `SpatialPolygonsDataFrame` from package **sp** (Bivand *et al.*, 2013) is provided additionally to the S3 object from `NER_Trafo`. As in **emdi** (Kreutzmann *et al.*, 2019), the same polygon data showing Austrian districts is available within **saeTrafo**, so that it is possible to visualize the estimates on a map. The `load_shapeaustria` function loads this map and the `map_plot` function offers various options for the users. This function directly supplies settings for the graphical representation (`color`, `scale_points`, and `guide`), outputs the processed data (`return_data`), and enables options to customize the map with the help of **ggplot2** (Wickham, 2016). If the domain IDs within the `SpatialPolygonsDataFrame` and the S3 object differ, `map_tab` enables the entry of a `data.frame` for the assignment of the domain IDs.

```
R> load_shapeaustria()

R> map_plot(NER_model, map_obj = shape_austria_dis,
+         map_dom_id = "PB")
```

The map in Figure 4 shows the mean equivalized household income for all 94 Austrian districts produced by the SAE methods explained above. Smaller values are mostly in rural districts (like Zell am See with the lowest value of 10469.93€) and higher mean equivalized household incomes appear in more urban districts.

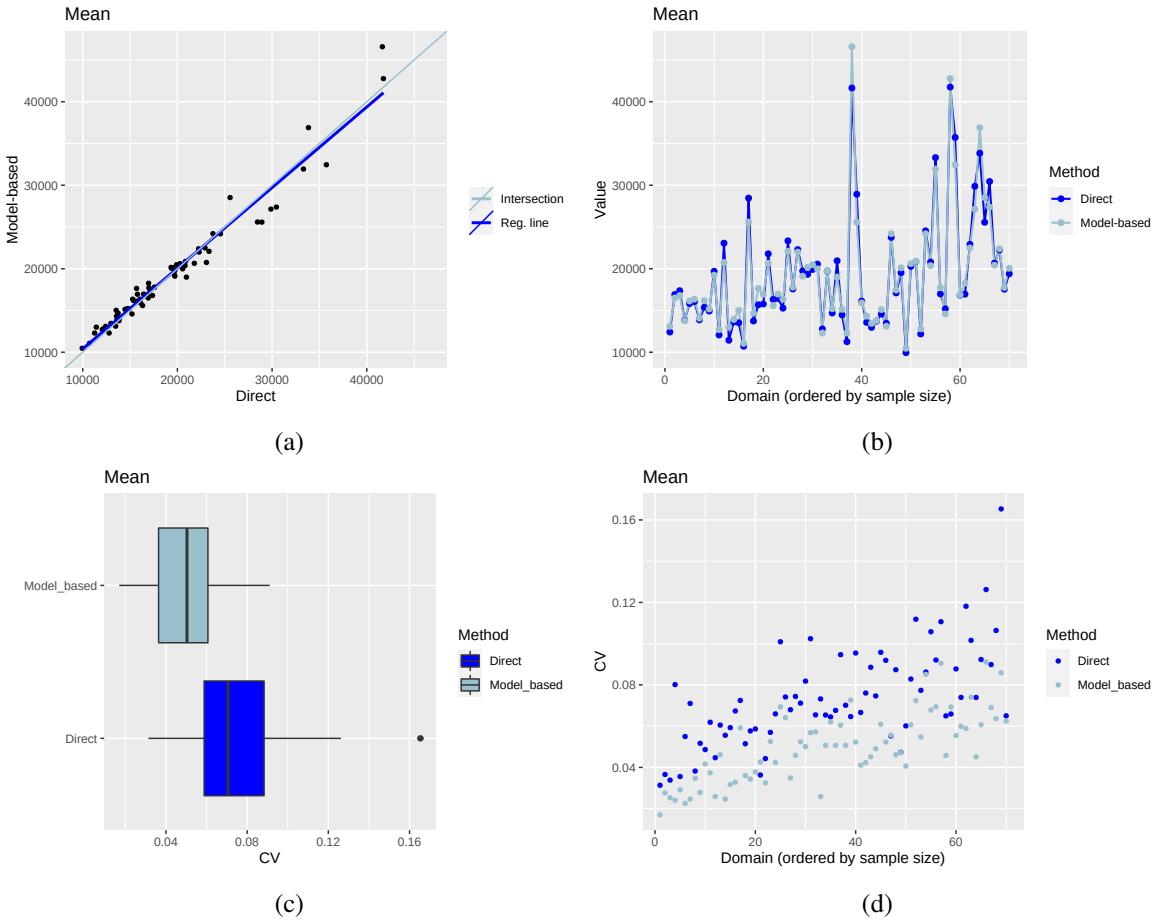

Figure 3: Plots for comparison to direct estimates from generic function `compare_plot` for the NER model: scatter plot (a), line plots with estimates ordered by domain-specific sample size (b), boxplots to compare CV for both estimators (c), and scatter plot for CV estimates ordered by domain-specific sample size (d).

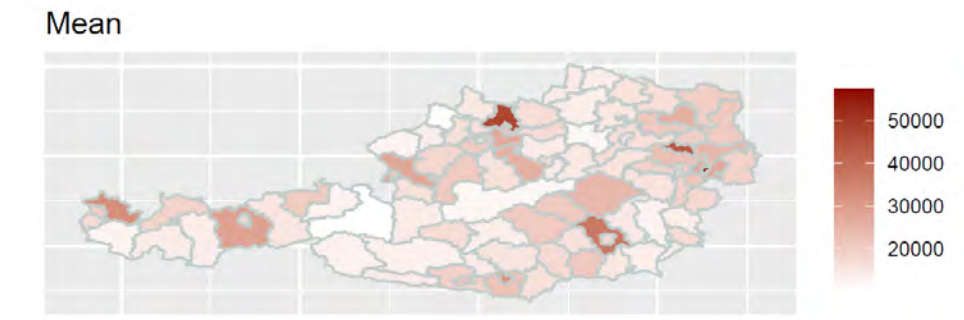

Figure 4: Map with Austrian districts showing their small area means for the equivalized household income from function `map_plot`.

Exporting the results and most important model information In addition to the evaluation and visualization of the point estimates (and uncertainty estimates), the package enables the export to other software. **saeTrafo** offers direct and user-friendly export of the estimates and the information from the `summary` function on the **saeTrafo** object to the software Excel.

```
R> write.excel(NER_model, file = "excel_output.xlsx",
+         CV = TRUE)
```

In addition, the export to `OpenDocument` format is also supported.

```
R> write.ods(NER_model, file = "excel_output.xlsx", CV = TRUE)
```

In both functions it can be specified if the CVs and MSEs should also be exported. If `split` is set to `TRUE`, the point estimators, MSEs and CVs are saved in separate worksheets, respectively separate documents. The created files are stored in the working directory.

Further generic functions Besides the generic functions already presented in detail, **saeTrafo** offers further generics: the function `estimators` is convenient to get point, MSE and CV estimates. In addition, the widely known functions `as.data.frame`, `as.matrix`, `head`, `print`, `subset`, and `tail` can be applied to the S3 object created with `estimators`. The `print` function returns the most important model information. To facilitate the comparison between SAE estimators, the generic function `compare_pred` exists and creates a data set with point or MSE estimators of both objects. To also enable comparisons with other SAE methodology, an **emdi** object can be entered.

To further increase user-friendliness, well-known, and widely used generic functions from the **stats** package can be used with **saeTrafo**. Thus, the following functions can be applied to the S3 object of `NER_Trafo`: `coef`, `confint`, `family`, `fitted`, `formula`, `logLik`, `nobs`, `predict`, `residuals`, `sigma`, `terms`, and `vcov`.

Since the linear mixed models used are calculated with the **nlme** package (Pinheiro *et al.*, 2022), the following generic functions for **nlme** objects are available for the S3 object of **saeTrafo**: `fixef`, `getData`, `getGroups`, `getGroupsFormula`, `getResponse`, `getVarCov`, `intervals`, and `ranef`.

5 Conclusion

The main focus of **saeTrafo** is to make the new methodology by Würz *et al.* (2022) publicly available. This methodology resolves the problem of not having access to individual population data while using transformations in the context of unit-level small area models. This method and its uncertainty estimation are supplied by the function `NER_Trafo`. In addition, the package provides the following methods: the well-known estimator by Battese *et al.* (1988), the bias-corrected estimator from Molina and Martín (2018) using population micro-data, and a first-order bias-corrected estimator using aggregated population data. An advantage of this function is the appropriate and automatic selection of small area methodology under different possible data inputs and transformations (none, log, and data-driven log-shift transformation). **saeTrafo** guarantees user-friendliness by providing all methods and their respective MSE (including parallelization options) within the `NER_Trafo` function. For this S3 object, a variety of generic functions are offered. They automate the creation of important plots for model diagnostics and the assessment of the estimator's quality. Furthermore, options for visualizing the estimates on maps and the export of estimators are provided. Further generic functionalities increase the user-friendliness.

This last paragraph outlines possible new features of **saeTrafo** for future releases: The choice between different methodologies to estimate the MSE will increase user-friendliness. For the estimator $\overline{\hat{Y}}_i^{\text{trans,bc}}$, Molina and Martín (2018) propose an analytical MSE in addition to the bootstrap version already supplied in **saeTrafo**. Further releases would profit by including this version. Moreover, **saeTrafo** offers the MSE of Prasad and Rao (1990) for the classical NER model. Further MSE estimating options are desirable. To have a MSE for the first-order bias-corrected estimator (trans, bc-naive-agg), theoretical

research is first necessary. Including alternative SAE methods such as the method of Li *et al.* (2019) will increase the flexibility of the package. Overall, the **saeTrafo** software package is written in such a way that this can be easily extended with other small area model classes. For long-term future versions, this is aspired.

Acknowledgements

Würz gratefully acknowledges support by a scholarship of Studienstiftung des deutschen Volkes.

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.

Alfons, A., Templ, M. and Filzmoser, P. (2010) An object-oriented framework for statistical simulation: the R package **simFrame**. *Journal of Statistical Software*, **37**, 1–36.

Barton, K. (2018) ***MuMIn**: Multi-Model Inference*. URL `https://CRAN.R-project.org/package=MuMIn`. R package version 1.40.4.

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.

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

Boonstra, H. J. (2021) ***mcmcsae**: Markov Chain Monte Carlo Small Area Estimation*. URL `https://CRAN.R-project.org/package=mcmcsae`. R package version 0.7.0.

Boonstra, H. J. (2022) ***hbsae**: Hierarchical Bayesian Small Area Estimation*. URL `https://CRAN.R-project.org/package=hbsae`. R package version 1.2.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodological)*, **26**, 211–252.

Breidenbach, J. (2018) ***JoSAE**: Unit-Level and Area-Level Small Area Estimation*. URL `https://CRAN.R-project.org/package=JoSAE`. R package version 0.3.3.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001) Evaluation of small area estimation methods - an application to unemployment estimates from the UK LFS. In *Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: A Methodological Perspective*. Statistics Canada.

Chambers, J. M. and Hastie, T. J. (1992) *Statistical Models in S*. London: Chapman & Hall.

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

Duong, T. (2022) ***ks**: Kernel Smoothing*. URL `https://CRAN.R-project.org/package=ks`. R package version 1.13.4.

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

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.

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.

Hagenaars, A., de Vos, K. and Zaidi, M. A. (1994) *Poverty Statistics in the Late 1980s: Research Based on Mirco-data*. Luxembourg: Office for the Official Publications of the European Communities.

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

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.

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

Lumley, T. (2004) Analysis of complex survey samples. *Journal of Statistical Software*, **9**, 1–19.

Maechler, M., Stahel, W., Ruckstuhl, A., Keller, C., Hauser, A., Buser, C., Gygax, L., Venables, B., Plate, T., Flückiger, I., Wolbers, M., Keller, M., Dudoit, S., Fridlyand, J., Snow, G., Nielsen, H. A., Carey, V., Bolker, B., Grosjean, P., Ibanez, F., Savi, C., Geyer, C. and Oehlschlägel, J. (2021) ***sfsmisc**: Utilities from 'Seminar fuer Statistik' ETH Zurich*. URL `https://CRAN.R-project.org/package=sfsmisc`. R package version 1.1-12.

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.

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

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

Pinheiro, J., Bates, D. and R Core Team (2022) ***nlme**: Linear and Nonlinear Mixed Effects Models*. URL `https://CRAN.R-project.org/package=nlme`. R package version 3.1-155.

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

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

R Core Team (2022) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. URL `\url{https://www.R-project.org}`.

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.

Schoch, T. (2014) ***rsae**: Robust Small Area Estimation*. URL `https://CRAN.R-project.org/package=rsae`. R package version 0.1-5.

Scott, D. W. (2015) *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken: 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.

Sugasawa, S. (2016) ***rhnerm**: Random Heteroscedastic Nested Error Regression*. URL `https://CRAN.R-project.org/package=rhnerm`. R package version 1.1.

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.

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

Würz, N., Schmid, T. and Tzavidis, N. (2022) Estimating regional income indicators under transformations and access to limited population auxiliary information. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, forthcoming.

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

Yang, Z. (2006) A modified family of power transformations. *Economics Letters*, **92**, 14–19.