Freie Universität Berlin

Data-Driven Transformations In Small Area Estimation

Natalia Rojas-Perilla
Sören Pannier
Timo Schmid
Nikos Tzavidis

School of Business & Economics
Discussion Paper

Economics

2017/30

DATA-DRIVEN TRANSFORMATIONS IN SMALL AREA ESTIMATION

By Natalia Rojas-Perilla [*], Sören Pannier [*], Timo Schmid [*] and Nikos Tzavidis [†]

Institute of Statistics and Econometrics, Freie Universität Berlin, Germany[*] and Southampton Statistical Sciences Research Institute, University of Southampton, UK [†]

Small area models typically depend on the validity of model assumptions. For example, a commonly used version of the Empirical Best Predictor relies on the Gaussian assumptions of the error terms of the linear mixed model, a feature rarely observed in applications with real data. The present paper proposes to tackle the potential lack of validity of the model assumptions by using data-driven scaled transformations as opposed to ad-hoc chosen transformations. Different types of transformations are explored, the estimation of the transformation parameters is studied in detail under a linear mixed model and transformations are used in small area prediction of linear and non-linear parameters. The use of scaled transformations is crucial as it allows for fitting the linear mixed model with standard software and hence it simplifies the work of the data analyst. Mean squared error estimation that accounts for the uncertainty due to the estimation of the transformation parameters is explored using parametric and semi-parametric (wild) bootstrap. The proposed methods are illustrated using real survey and census data for estimating income deprivation parameters for municipalities in the Mexican state of Guerrero. Extensive simulation studies and the results from the application show that using carefully selected, data driven transformations can improve small area estimation.

1. Introduction. Model-based methods for small area estimation (SAE) are now widely used in practice for producing reliable estimates of linear and non-linear indicators for areas/domains with small sample sizes. Examples of indicators that are estimated by using model-based methods include poverty (income deprivation) and inequality measures such as the head count ratio, the poverty gap and the Gini coefficient. Two popular small area methods in this case are the empirical best predictor (EBP), proposed by Molina and Rao [1] and the World Bank method, proposed by Elbers et al. [2]. Both approaches are based on the use of unit-level linear mixed regression models.

Keywords and phrases: Small area estimation, linear mixed regression model, MSE estimation, data-driven transformations, poverty mapping, maximum likelihood theory.

Although estimation of complex indicators can be also implemented with area-level models [3, 4], we pay particular attention to unit-level models in this paper. Focusing on the EBP method, we note that the corresponding theory only exists under specific distributions. In the original paper, Molina and Rao [1] assumed that the error terms of the linear mixed regression model follow a Gaussian distribution. Molina and Rao [5] noted that, if the model error terms significantly deviate from normality, the EBP estimator can be biased. What are the options available to the data analyst when the normality assumptions are not met? A first option is to formulate the EBP under alternative and more flexible parametric assumptions. Graf et al. [6] study an EBP method under the generalized beta distribution of the second kind (GB2), whereas Diallo and Rao [7] propose the use of skewed-normal distributions in applications with income data. One complication with using the EBP under alternative parametric distributions is that new tools for estimation must be developed and training for the data analyst is needed. In addition, misspecification of the model assumptions is still possible. The second option when the Gaussian assumptions are not satisfied is to use a methodology that minimizes the use of parametric assumptions. Weidenhammer et al. [8] recently proposed a method that aims at estimating the quantiles of the empirical distribution function of the data. The estimation of the quantiles is facilitated by a nested error regression model using the asymmetric Laplace distribution for the unit-level error terms as a working assumption. The estimation of the random effects can be made completely non-parametric by using a discrete mixture proposed by Marino et al. [9]. However, this method does not necessarily lead to an empirical best predictor, while implementation also requires the development of new estimation and inference tools. A third option, and the one we study in this paper, is to find an appropriate transformation such that the model assumptions (in this paper the Gaussian assumptions of the EBP method) hold. The aim is to find transformations that (a) are data-driven and optimal according to some criterion and (b) can be implemented by using standard software. To the best of our knowledge, the use and choice of transformations has not been studied extensively or it has been studied in fairly ad-hoc manner. Elbers et al. [2] and Molina and Rao [1] suggested the use of logarithmic-type transformations for income data. However, is such a transformation the most appropriate choice? Can alternative transformations offer improved estimation? In order to answer these research questions, the paper investigates data-driven transformations for small area estimation.

The choice of transformations when modelling income-type outcomes - as is the case with poverty mapping applications - presents different challenges.

Transformations should be suitable for dealing with unimodal, leptokurtic and positively skewed data that may include zero and negative values. Besides the logarithmic transformation and its modifications (e.g. the log-shift transformation) a popular family of data-driven transformations that includes the logarithmic one as a special case is the Box-Cox [10] family. Since the Box-Cox transformation is not defined for negative values, when negative values are present, the data must be shifted to the positive range. Another difficulty with the use of the Box-Cox transformation is the truncation on the transformation parameter described later in Section 4. A solution to this problem can be offered by the use of the dual power transformation. Although very rich literature on the use of transformations exists see for example, Manly [11], John and Draper [12], Bickel and Doksum [13] and Yeo and Johnson [14], among others. In this paper we focus on three types of transformations, namely log-shift, Box-Cox and dual power transformations.

In addition to selecting the type of transformation, estimating the transformation parameter adds another layer of complexity. To the best of our knowledge the use of transformations in recent applications of small area estimation has employed visual residual diagnostics for finding a suitable transformation parameter. In this paper we propose a structured, data-driven approach for estimating the transformation parameter. In particular, we introduce maximum likelihood and residual maximum likelihood methods for estimating the transformation parameter under a linear mixed regression model. Alternative estimation approaches based on the minimization of distances [15, 16] and on the minimization of the skewness [17] are also discussed. The use of scaled transformations such that standard software can be employed for estimation and inference is another important aspect of what we propose in this paper.

We study how the performance of the EBP method is affected by departures from normality and how data-driven transformations can assist with improving the validity of the model assumptions and estimation. Emphasis is given to the estimation of poverty and inequality indicators due to their important socio-economic relevance and policy impact. We further study whether the impact of departures from Gaussian assumptions is different depending on the target of estimation. For example, departures from normality may have less impact on estimates of median income compared to estimates of the Gini coefficient. The estimation for the latter indicator heavily depends on the entire distribution of the data. A parametric bootstrap for mean squared error (MSE) estimation under transformation is studied and a wild-type bootstrap that may offer protection in the presence of departures from the Gaussian assumptions after transformations is also proposed.

The paper is structured as follows. The details of the EBP approach are introduced in Section 2. Section 3 presents the survey data we use in this paper and makes the case, via the use of residual diagnostics, for using transformations. In Section 4 a number of possible transformations are introduced and extended for their use with model-based SAE methods under a linear mixed regression model. This section includes the theoretical details about the choice of an appropriate scale and estimation of the transformation parameter. MSE estimation is discussed in Section 5. The proposed methods are applied to data from Guerrero in Mexico for estimating a range of deprivation and inequality indicators and corresponding estimates of uncertainty in Section 6. In Section 7 the proposed methods are further evaluated via the use of model-based simulation under realistic scenarios for income data. Section 8 summarizes the main findings and outlines further research.

2. Empirical best prediction with applications for estimating poverty and inequality. Let U denote a finite population of size N partitioned into D areas or domains (representing the small areas) $U_1, U_2, \ldots, U_D$ of sizes $N_1, \ldots, N_D$, where $i = 1, \ldots, D$ refers to the ith area. Let y_{ij} be the target variable defined for the jth individual belonging to the ith area, with $j = 1, \ldots, N_i$. Denote by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^T$ the design matrix containing p explanatory variables and define by s as the set of sample units, with s_i the in-sample units in area i and by r be the set of non-sampled units, with r_i the out-of-sample units in area i. Let n_i denote the sample size in area i with $n = \sum_{i=1}^{D} n_i$. Hence, we define by $\mathbf{y}_i$ a vector with population elements of the target outcome for area i partitioned as $\mathbf{y}_i^T = \left(\mathbf{y}_{is}^T, \mathbf{y}_{ir}^T\right)$, where $\mathbf{y}_{is}$ and $\mathbf{y}_{ir}$ denote the sample elements s and the out-of-sample elements r in area i respectively. Let us now describe in more detail the EBP approach by Molina and Rao [1], which is the methodology we focus on in this paper. Under this approach census predictions of the target outcome are generated by using the conditional predictive distribution of the out-of-sample data given the sample data. The point of departure is the standard parametric unit-level linear mixed regression model, which is also known as the unit-level nested error regression model and for simplicity, is called in this paper as the linear mixed regression model. This is defined by Battese et al. [18] as:

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i + e_{ij}, \quad u_i \overset{iid}{\sim} N(0, \sigma_u^2) \quad \text{and} \quad e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2), \tag{2.1}$$

where u_i, the area-specific random effects, and e_{ij}, the unit-level errors, are assumed to be independent. Assuming normality for the unit-level error and the area random effects, the conditional distribution of the out-of-sample

data given the sample data are also normal. A Monte Carlo approach is used to obtain a numerically efficient approximation to the expected value of this conditional distribution as follows:

1. Use the sample data to obtain $\hat{\boldsymbol{\theta}}^T = (\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$ and $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.
2. For $l = 1, \ldots, L$:
   (a) Generate $v_i^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_u^2(1-\hat{\gamma}_i)\big)$ and $e_{ij}^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_e^2\big)$ and obtain a pseudo-population of the target variable by:
   $$y_{ij}{}^{(l)} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_i + v_i^{(l)} + e_{ij}^{(l)},$$
   where the predicted random effect $\hat{u}_i$ is defined as $\hat{u}_i = E(u_i|\mathbf{y}_{is})$.
   (b) Calculate the indicator of interest $I_i^{(l)}$ in each area.
3. Finally, take the mean over the L Monte Carlo runs in each area to obtain a point estimate of the indicator of interest:
   $$\hat{I}_i^{EBP} = \frac{1}{L}\sum_{l=1}^{L} I_i^{(l)}.$$

As is common in real applications, some areas are out-of-sample. For those areas, we cannot estimate a random effect, and hence the corresponding random effect is set equal to zero. Synthetic values of the outcome for the out-of-sample areas are then generated under the linear mixed regression model as follows:

$$y_{ij}{}^{(l)} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_i^{(l)} + e_{ij}^{(l)},$$

with $u_i^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_u^2\big)$ and $e_{ij}^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_e^2\big)$. Finally, a parametric bootstrap - under the assumed model - is used for the MSE estimation. This is discussed in some detail in Section 5. As we mentioned in Section 1, the EBP method is applicable only by making specific parametric assumptions about the distribution of the error terms of the linear mixed regression model that allow the conditional distribution $\mathbf{y}_r|\mathbf{y}_s$ to be obtained. From the point of view of the analyst, the easiest (but not the only option) is to assume normality since standard software for fitting the linear mixed regression model is available in this case. However, in applications, which involve modelling an income-type outcome, assuming normality is unrealistic. In this paper we use such an outcome for estimating indicators such as the head count ratio [19], the income quintile share ratio [20] and the Gini coefficient [21]. If our primary concern is to develop a methodology that can easily be used in practice, finding appropriate data transformations to normality is of paramount importance.

3. The Guerrero case study: Data source and initial analysis. In this section, we describe the data sources used in the application and provide a motivation for the use of transformations.

The data used in this paper come from Mexico, which has one of the largest economies in Latin America and is still among the most unequal countries in the world, according to the World Bank. For tailored policies against deprivation, it is necessary to know a detailed description of the spatial distribution of inequality and income deprivation. According to the general social development law in Mexico, the National Institute of Statistics and Geography (INEGI) has to provide measures at the national, state and municipal-level. For carrying out the analysis in this paper, the statistical and geographical information was provided by INEGI through the Household Income and Expenditure Survey (ENIGH) 2010 and the National Population and Housing Census of 2010. Looking in more detail at the data available and their geographic coverage, Mexico is divided into 32 federal entities (states). The state Guerrero has been considered by the World Bank to be one of the entities that mostly contributes to inequality in Mexico, presenting a high inequality in human development. Additionally, according to the United Nations Development Programme (UNDP), this region presents one of the highest rates of poverty and lack of infrastructural development. Guerrero is made up of 81 administrative divisions, known as municipalities. From the 81 municipalities in Guerrero, only 40 are in-sample, leaving the other 41 out-of-sample. Furthermore, there are 1611 households in the 40 in-sample municipalities. Table 1 shows a summary of the sample sizes for these municipalities, in which the maximum sample size in a municipality is 511, whereby the minimum is only 9 and the median is 24 households.

TABLE 1
Sample sizes of the municipalities available in survey data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Sample	9.00	17.00	24.00	40.27	36.00	511.00

The survey and census data include a large number of socio-demographic variables, which are common and are measured similarly in both data sources. The total household per capita income from work, denoted by *hciw* is an example of a variable which is available in the survey but not in the census. the variable *hciw* is used in this paper as the variable that best approximates the living standard in households in Guerrero. Therefore, the working model used in this paper defines as the outcome variable, *hciw*, which is measured in Mexican pesos. Available socio-demographic variables of the households are utilized as explanatory variables. The underlying linear mixed regres-

sion model (2.1) of the EBP is characterized by having two levels, in which households are grouped by municipalities. The variables available in the survey and census data, which are identified by using the Bayesian information criterion (BIC) as good predictors of *hciw*, are described in Table 2.

TABLE 2
Description of the explanatory variables used in the working model

Determinant	Variable
Occupation	1) Indicator if the head of household and the spouse are employed
	2) Type of household occupation
	3) Total number of employees older than 14 years in a household
	4) Percentage of employees older than 14 years in a household
Sources of income	5) Indicator of a household receiving remittances
Socioeconomic level	6) Availability of assets in the household
	7) Total number of goods in the household
Education	8) Average standardized years of schooling (by age and sex) within the household relative to the population

The next step after the identification of a possible set of covariates is assessing the predictive power of the model. Nakagawa and Schielzeth [22] propose the use of two coefficients of determination suitable for generalized mixed-effects regression models: (a) the marginal R^2_m, which is a measure for the variance explained by fixed effects and (b) the conditional R^2_c, which measures the variance explained by both, the fixed and random effects. Without using any transformation, these measures are both around 34% and the corresponding intraclass correlation (ICC) under the model is 0.02.

In order to explore the validity of the Gaussian assumptions underlying the linear mixed regression model, it is appropriate to perform normality tests and some residual diagnostics. The p-values of the Shapiro-Wilk (S-W) test statistic are equal to $2.2 \cdot 10^{-16}$ for the household-level and 0.002 for the municipal-level. These results indicate that the null hypothesis of normality for both terms has to be rejected. Additionally, Figure 1 presents the Normal probability quantile-quantile (Q-Q) plots for household-level and municipal-level residuals. As expected, in the case of using the non-transformed *hciw* variable, the shape of the Q-Q plots is clearly different from what would be expected under normality. In addition, the analysis of skewness and kurtosis for both error terms is also informative. The skewness and kurtosis for a Normal distribution are equal to zero and three, respectively. The results for the skewness and kurtosis on the household-level are equal to 7.980 and 110.700, and on the municipal-level equal to 1.298 and 5.596. These results indicate severe departures from Gaussian assumptions when modelling the non-transformed income data.

As mentioned before, one solution to tackling this problem is based on

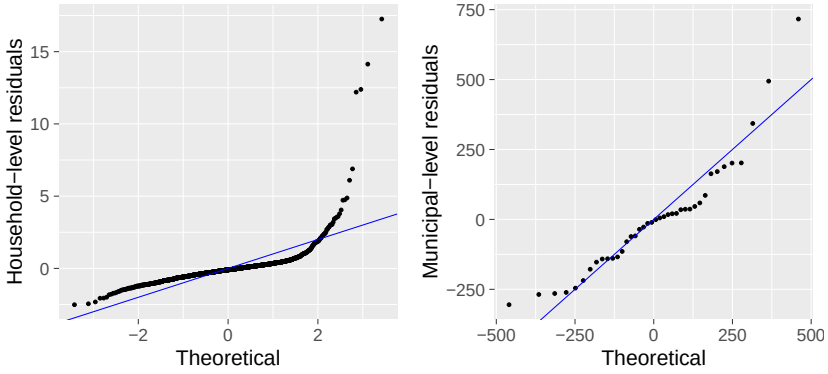

FIG 1. *Q-Q plots of the household- and municipal-level error terms*

using transformations before opting for applying more complex methodologies. The challenge in this case is finding an appropriate transformation for the data, such that the normality assumptions of the underlying model are met.

4. Use of transformations. As previously seen in Sections 2 and 3, the EBP method relies on strong distributional assumptions, which are hardly fulfilled in applications with income data. To ensure normality, it is very common to use a one-to-one transformation $T(y_{ij}) = y_{ij}^*$ of the target variable $\boldsymbol{y}$ [23]. The application of the natural logarithmic transformation, which is a popular choice for income data, leads in many cases from highly right-skewed to more symmetric distributions. This is the most frequently used transformation in different research fields for dealing with non-normality due to its simplicity and because no additional training for users is required. However, can an alternative transformation with data-driven transformation parameter λ, $T_\lambda(y_{ij}) = y_{ij}^*(\lambda)$, possibly offer small area estimates with improved precision?

The structure of the Section is as follows: Firstly, in Section 4.1 we introduce briefly the EBP approach with data-driven transformations. Secondly, in Section 4.2 we propose likelihood-based approaches for estimating the adaptive transformation parameter in general and discuss three particular subcases - namely log-shift, Box-Cox and dual power transformations - in detail. Thirdly, in Section 4.3 we discuss alternative approaches for estimating the transformation parameter.

4.1. *EBP approach under transformations.* In order to apply the EBP method by using transformations, the linear mixed regression model given in 2.1 is re-defined as follows:

$$y_{ij}^*(\lambda) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i + e_{ij}, \quad u_i \overset{iid}{\sim} N\big(0, \sigma_u^2\big) \text{ and } e_{ij} \overset{iid}{\sim} N\big(0, \sigma_e^2\big). \tag{4.1}$$

Consequently, the EBP approach under transformations can be re-written as follows:

1. Select a transformation and obtain $T_\lambda(y_{ij}) = y_{ij}^*(\lambda)$.
2. Use the transformed sample data to obtain $\hat{\boldsymbol{\theta}}^T = \big(\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2\big)$ and calculate the weighting factors, $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.
3. For $l = 1, \ldots, L$:
   (a) Generate $v_i^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_u^2(1-\hat{\gamma}_i)\big)$ and $e_{ij}^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_e^2\big)$ and obtain a pseudo-population of the target variable by:
   $$y_{ij}{}^{*(l)} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \hat{u}_i + v_i^{(l)} + e_{ij}^{(l)}.$$
   (b) Back-transform $y_{ij}{}^{*(l)}$ to the original scale $y_{ij}^{(l)} = T_\lambda^{-1}\left(y_{ij}^{*(l)}\right)$.
   (c) Calculate the indicator of interest $I_i^{(l)}$ in each area.
4. Finally, take the mean over the L Monte Carlo generations in each region to obtain an approximation of the indicator of interest:
$$\hat{I}_i^{EBP} = \frac{1}{L}\sum_{l=1}^{L} I_i^{(l)}.$$

4.2. *Likelihood-based approach for estimating* λ. For the estimation of the transformation parameter λ, the linear mixed regression model defined in 4.1 is used. Assume that the transformed vectors $\boldsymbol{y}_i^*$ are independent and normally distributed for some unknown λ as follows:

$$\boldsymbol{y}_i^*(\lambda) \sim N(\boldsymbol{\mu}_i, \mathbf{V}_i) \quad \text{for} \quad i = 1, \ldots, D,$$

where

$$\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta} \quad \text{and} \quad \mathbf{V}_i = \sigma_u^2\mathbf{1}_{N_i}\mathbf{1}_{N_i}^{'} + \sigma_e^2\mathbf{I}_{N_i},$$

with $\mathbf{1}_{N_i}$ a column vector of ones of size N_i and $\boldsymbol{I}_{N_i}$ the $N_i \times N_i$ identity matrix, the vector of unknown model parameters is $\boldsymbol{\theta}^T = (\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2, \lambda)$. The

log-likelihood function under the model is defined as follows:

$$\begin{aligned} l_{\mathrm{ML}}(\boldsymbol{y}^*, \lambda|\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]. \end{aligned}$$

The log-likelihood function in relation to the original observations is obtained by multiplying the normal density by $J(\lambda, \boldsymbol{y})$, the Jacobian of the transformation from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^*(\lambda)$. The Jacobian is defined as $\prod_{i=1}^{D}\prod_{j=1}^{n}\left|\frac{dy_{ij}^*(\lambda)}{dy_{ij}}\right|$ and is incorporated as follows:

$$\begin{aligned} l_{\mathrm{ML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]\ J(\lambda, \boldsymbol{y}). \end{aligned}$$

The maximization of $l_{\mathrm{ML}}(\boldsymbol{\theta})$ produces maximum likelihood (ML) estimates of the unknown parameters $\boldsymbol{\theta}$. However, in the theory of linear mixed regression models, when interest focuses on accurate estimators of the variance components, restricted maximum likelihood (REML) theory is recommended [24]. The REML function, in which the maximum possible number of linear independent contrasts is $n-p$ [25], does not depend on $\boldsymbol{\beta}$ is defined as follows:

$$\begin{aligned} l_{\mathrm{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]\ J(\lambda, \boldsymbol{y}). \end{aligned} \tag{4.2}$$

To take advantage of procedures for estimating λ already computationally implemented we convert a transformation in a so-called, scaled transformation by $\frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{\frac{1}{n}}} = z_{ij}^*(\lambda)$. The scaled transformations are conditioned on the

Jacobian, which is equal to 1. Therefore, the residual log-likelihood function in terms of the original observations is that of the linear mixed regression model:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{z}^*, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{z}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{z}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]. \end{aligned}$$

Although the theory is applicable to data-driven transformations in general, we focus on three types of transformations, namely log-shift, Box-Cox and dual power transformations as particular subcases. The first data-driven transformation is called log-shift [26]. It is extending the logarithmic function by including the transformation parameter λ as follows:

$$y_{ij}^*(\lambda) = \log(y_{ij} + \lambda).$$

When $\lambda = 0$, a logarithmic transformation is obtained. The second type is the Box-Cox [10] transformation, which is defined by:

$$y_{ij}^*(\lambda) = \begin{cases} \frac{(y_{ij}+s)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(y_{ij}+s) & \text{if } \lambda = 0, \end{cases}$$

where s denotes a fixed parameter such that $y_{ij} + s > 0$ to enable the use of the Box-Cox transformation. When $\lambda = 0$, the logarithmic transformation is then a special case of this family and if $\lambda = 1$, the data are only shifted. One difficulty with the Box-Cox type transformations is the long-standing truncation, i.e. $y_{ij}^*(\lambda)$ is bounded, from below by $\frac{1}{\lambda}$ if $\lambda > 0$ and from above by $\frac{-1}{\lambda}$ if $\lambda < 0$. This is the key motivation for the third type of transformation. The dual power transformation introduced by Yang [27] is defined as follows:

$$y_{ij}^*(\lambda) = \begin{cases} \frac{(y_{ij}+s)^\lambda - (y_{ij}+s)^{-\lambda}}{2\lambda} & \text{if } \lambda > 0, \\ \log(y_{ij}+s) & \text{if } \lambda = 0, \end{cases}$$

where s is defined as in the case of Box-Cox transformations.

The corresponding Jacobian used in Equation 4.2 and scaled versions of the log-shift, Box-Cox and dual power transformations are presented in Table 3. For more details we refer to the proofs in Appendix A.

TABLE 3
Jacobian and scaled data-driven transformations for log-shift, Box-Cox and dual

Transformation	Jacobian J	Scaled transformation $z^*_{ij}(\lambda)$
Log-Shift	$\prod_{i=1}^{D}\prod_{j=1}^{n} y_{ij}^{-1}$	$J^{\frac{-1}{n}} \log(y_{ij}+\lambda)$
Box-Cox	$\prod_{i=1}^{D}\prod_{j=1}^{n} y_{ij}^{\lambda-1}$	$J^{\frac{-1}{n}} \frac{(y_{ij}+s)^{\lambda}-1}{\lambda}$, if $\lambda \neq 0$ $J^{\frac{-1}{n}} \log(y_{ij}+s)$, if $\lambda = 0$
Dual	$\frac{\prod_{i=1}^{D}(\prod_{j=1}^{n} y_{ij}^{\lambda-1}+y_{ij}^{-\lambda-1})}{2}$	$J^{\frac{-1}{n}} \frac{(y_{ij}+s)^{\lambda}-(y_{ij}+s)^{-\lambda}}{2\lambda}$ if $\lambda \neq 0$ $J^{\frac{-1}{n}} \log(y_{ij}+s)$ if $\lambda = 0$

4.3. *Alternative approaches for estimating* λ. According to [28] and [29], different considerations about the likelihood-based method for estimating the transformation parameter might be taken into account. Firstly, the ML and REML approaches introduced in Subsection 4.2 rely on parametric assumptions that may be influenced by outliers in the data. Secondly, the selected transformation needs to be differentiable in order to guarantee the existence of the Jacobian J of the transformation. A non-parametric alternative for estimating λ is to focus directly on optimizing the form of the distribution of the error terms, for instance, by three- and four-moment optimization strategies and by estimators based on divergence optimization.

As the kurtosis and skewness are crucial features for defining the shape of a normal distribution, a proximity measure may be minimized in order to achieve residuals whose skewness or kurtosis is as close as possible to zero and three, respectively. In general, skewness is considered more disturbing than kurtosis. Therefore, minimizing skewness is an approach already considered in literature in linear models [30]. In the context of linear mixed regression models, an additional problem arises as there are two independent error terms to be considered. We propose a pooled skewness approach ensuring by a weight w that the larger the error term variance is, the more importance will have its skewness in the optimization. Let S_{e_λ} and S_{u_λ} be the skewness and $\sigma^2_{e_\lambda}$ and $\sigma^2_{u_\lambda}$ be the variance of the unit-level error terms e_{ij} and the random area-specific effects u_i of the linear mixed regression model, respectively. In this case, the index λ is incorporated to emphasize that the skewness and variances depend on the transformation parameter. The functional form of the estimation criteria is defined as follows:

$$\hat{\lambda}_{\text{skew}} = \underset{\lambda}{\text{argmin}} \, |S_{e_\lambda}|,$$
$$\hat{\lambda}_{\text{poolskew}} = \underset{\lambda}{\text{argmin}} \, \big(w|S_{e_\lambda}| + (1-w)|S_{u_\lambda}|\big),$$
$$\text{where } w = \frac{\hat{\sigma}^2_{e_\lambda}}{\hat{\sigma}^2_{u_\lambda} + \hat{\sigma}^2_{e_\lambda}}.$$

However, only considering skewness may ignore other properties of the distribution. Hence, a measure describing the distance between two distribution functions as a total might be another alternative. Two additional measures are the Kolmogorov-Smirnov (KS) [16] and the Cramér-von Mises (CvM) [15] distances:

$$\hat{\lambda}_{\text{KS}} = \underset{\lambda}{\text{argmin}} \, \sup \, |F_n(\cdot) - \Phi(\cdot)|,$$
$$\hat{\lambda}_{\text{CvM}} = \underset{\lambda}{\text{argmin}} \int_{-\infty}^{\infty} [F_n(\cdot) - \Phi(\cdot)]^2 \, \phi(\cdot),$$

where $F_n(\cdot)$ is the empirical cumulative distribution function estimated by using the normalized residuals, $\Phi(\cdot)$ is the distribution function of a standard normal distribution and $\phi(\cdot)$ its density. A generalization to a pooled measure as done for the skewness is applicable. The impact of alternative estimation approaches for λ will be compared in a model-based simulation study in Section 7.3.

5. MSE estimation under transformations. Molina and Rao [1] have already pointed out that the estimation of MSE is a challenging problem in the case of the EBP. They propose a parametric bootstrap procedure for the MSE estimation following the approach by González-Manteiga et al. [31]. In this section we propose two bootstrap schemes for estimating the MSE of the proposed transformed small area estimator we presented in Section 4. These bootstrap MSE estimators are extended to capture the additional uncertainty due to the estimation of the transformation parameter λ. The difference between the bootstrap schemes is the mechanism used for generating the bootstrap population. In particular, the first bootstrap approach generates parametrically bootstrap realisations of the random effects and error terms. In contrast, the second one is a semi-parametric wild bootstrap which protects against departures from the assumptions of the regression model in particular, those of the unit-level error term.

The steps of the proposed parametric bootstrap are as follows:

1. For $b = 1, ..., B$
   (a) Using the sample estimates $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}$ from the transformed data $T(y_{ij}) = y_{ij}^*$, generate $u_i^{(b)} \overset{iid}{\sim} N\left(0, \hat{\sigma}_u^2\right)$ and $e_{ij}^{(b)} \overset{iid}{\sim} N\left(0, \hat{\sigma}_e^2\right)$ and simulate a bootstrap super-population ${y_{ij}}^{*(b)} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_i^{(b)} + e_{ij}^{(b)}$.
   (b) Back-transform ${y_{ij}}^{*(b)}$ to the original scale $y_{ij}^{(b)} = T_\lambda^{-1}\left(y_{ij}^{*(b)}\right)$ and compute the population value of the indicator of interest $I_{i,b}$.
   (c) Extract the bootstrap sample in ${y_{ij}}^{(b)}$ and perform the EBP method, as described in Section 4.1. Note, as the back-transformed sample data are used, the transformation parameter λ is re-estimated in each bootstrap replication b.
   (d) Obtain $\hat{I}_{i,b}^{EBP}$.
2. $\widehat{MSE}\left(\hat{I}_i^{EBP}\right) = B^{-1} \sum_{b=1}^{B} \left(\hat{I}_{i,b}^{EBP} - I_{i,b}\right)^2$.

As mentioned before, the proposed parametric bootstrap accounts for the additional uncertainty due to the estimation of the transformation parameter. Although the use of an optimal transformation reduces the deviation from normality, there may still be *small* departures from normality, especially in the tails of the distributions of the unit-level error term affecting the MSE estimation based on the parametric bootstrap proposed above. To overcome this problem, we introduce a semi-parametric bootstrap approach which relies on a normality assumptions of the random effects, but generates the unit-level error terms by a non-parametric approach. In particular, the semi-parametric bootstrap extends the theory of wild bootstrap [32, 33] to linear mixed regression models. The proposed wild bootstrap scheme is described below:

1. Use the sample estimates $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}$ from the transformed data $T(y_{ij}) = y_{ij}^*$.
2. Calculate the sample residuals by $\hat{e}_{ij} = y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} - \hat{u}_i$.
3. Scale and center the residuals according to $\hat{\sigma}_e$, which are denoted by $\hat{\epsilon}_{ij}$.
4. For $b = 1, ..., B$
   (a) Generate $u_i^{(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$.
   (b) Calculate the linear predictor $\eta_{ij}^{(b)}$ by $\eta_{ij}^{(b)} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_i^{(b)}$.
   (c) Match the population $\eta_{ij}^{(b)}$ and the sample $\hat{\eta}_k = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_i \;\; (k \in n)$ by

$$\min_{k \in n} \left| \eta_{ij}^{(b)} - \hat{\eta}_k \right|$$

and define $\tilde{k}$ as the corresponding index.

(d) Generate weights w from a distribution satisfying the conditions in Feng et al. [34] where w is a simple two-point mass distribution with probabilities 0.5 at $w = 1$ and $w = -1$, respectively.

(e) Calculate the bootstrap population as $y_{ij}^{*(b)} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + u_i^{(b)} + w_k|\hat{\epsilon}_{\tilde{k}}^{(b)}|$.

(f) Transform ${y_{ij}}^{*(b)}$ to original scale and compute the population value $I_{i,b}$.

(g) Extract the bootstrap sample in ${y_{ij}}^{(b)}$ and perform the EBP method, as described in Section 4.

(h) Obtain $\hat{I}_{i,b}^{EBP}$.

5. $\widehat{MSE}_{wild}\left(\hat{I}_i^{EBP}\right) = B^{-1}\sum_{b=1}^{B}\left(\hat{I}_{i,b}^{EBP} - I_{i,b}\right)^2$.

As the residuals are generated by a non-parametric approach, we expect that the proposed wild bootstrap protects against departures from normality of the unit-level error term. We will compare the performance of both MSE estimators in a model-based simulation study in Section 7.

6. The Guerrero case study: Application of data-driven transformations. In this section, the benefits of using the proposed EBP approach with data-driven transformation for the estimation of deprivation and inequality indicators are illustrated in an application using the data from the ENIGH survey 2010 and the National Population and Housing Census 2010 introduced in Section 3. The aim is to estimate the head count ratio (HCR) and the poverty gap (PGAP) introduced by Foster et al. [19] as well as the income quintile share ratio (QSR) [20] for the 81 municipalities in Guerrero.

Before we focus on the state of Guerrero, we shortly illustrate the need for data-driven transformations in states in Mexico. Figure 2 represents the estimated data-driven Box-Cox transformation parameters (by REML) for each state in Mexico. These estimates vary between 0.13 and 0.37, showing the adaptive feature of data-driven transformations for each state in Mexico. Furthermore, we observe that a fixed logarithmic transformation is not suitable for any of the states.

6.1. *Model checking and residual diagnostics.* We already observed in Section 3 that the model assumptions of the linear mixed regression model are not met. We now discuss the use of the proposed data-driven transformations to adapt the underlying model. In particular, we focus on the three data-driven transformations presented in Section 4.2, denoted by *Log-Shift*,

FIG 2. *Estimated transformation parameters of the Box-Cox transformation in the different states of Mexico*

Box-Cox and *Dual* power transformations and their comparison to (a) a model that use a logarithmic transformation (*Log*) and (b) a model that uses the untransformed income variable (*No*).

To start with, Figure 3 provides a graphical representation of the REML maximization for the transformation parameter λ for log-shift, Box-Cox and dual power transformations. In this case the optimal λs are approximately equal to 68.16, 0.26 and 0.30, respectively (cf. Table 4). In order to analyze

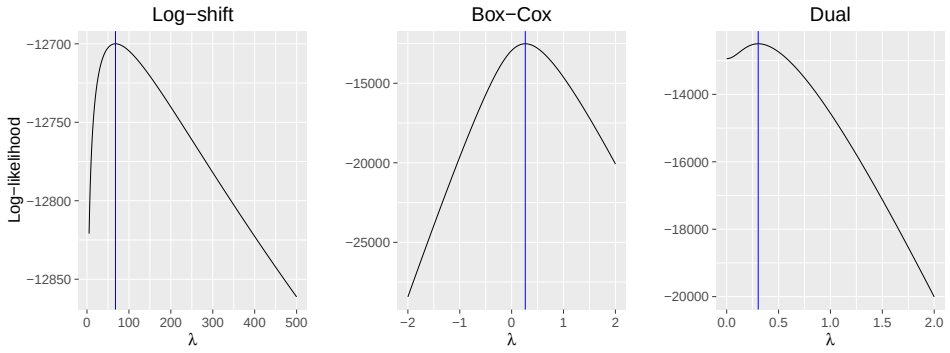

FIG 3. *Optimal transformation parameter λs for the log-shift, Box-Cox and dual power transformations*

whether the use of the transformations improves the predictive power of the model, Table 4 reports the percentage of variability explained for each model and its corresponding ICC. As the ICC is larger than 0 in all cases, the use of a linear mixed regression model seems to be appropriate. Using

the untransformed hciw data in the underlying model of EBP leads to a marginal (R^2_m) and conditional (R^2_c) coefficients of determination of 0.33 and 0.35, respectively. The use of a logarithmic transformation improves the predictive power of the model in terms of the conditional R^2_c but it loses in terms of marginal R^2_m. However, it can clearly be noted that the use of data-driven transformations increases the predictive power of the model.

TABLE 4
R^2_m, R^2_c, λ_s, and ICC for the working model under the different transformations

	R^2_m	R^2_c	λ	ICC
No	0.331	0.346	-	0.023
Log	0.263	0.416	-	0.207
Log-Shift	0.419	0.517	68.159	0.169
Box-Cox	0.439	0.517	0.263	0.140
Dual	0.443	0.517	0.304	0.132

Since the estimation of a linear mixed regression model and the EBP method depend on distributional assumptions, a detailed analysis of the Gaussian assumptions of the working models corresponding to each transformation is carried out. The results summarizing the skewness, kurtosis and S-W normality tests are presented in Table 5 and the Q-Q plots are presented in Figure 4. It should be noted, that at municipal-level, all three data-driven transformations perform similarly and yield good approximations to the normal distribution. In contrast, the household-level residuals show clear departures from normality, especially under the model with a fixed logarithmic transformation and without a transformation. The picture considerably improves for the data-driven transformations. The log-shift, Box-Cox and dual power transformations lead on average to very similar results in terms of skewness and kurtosis, with only small differences. We note that the log-shift transformation performs slightly better in terms of kurtosis, but not in terms of skewness compared to the Box-Cox and dual power transformation. These findings can also be supported by the Q-Q plots displayed in Figure 4. The data-driven transformations lead to similar Q-Q plots with more symmetrical and less extreme tails compared to the fixed log transformation. We observe only minor differences between the three data-driven transformations. For instance, it seems that the Box-Cox and dual transformations slightly suffer from some outliers on the right tail for the household-level residuals. Overall, it appears that the proposed data-driven transformations improve the predictive power of the model and clearly give better approximations to the underlying model assumptions of the linear mixed regression model compared to using a fixed logarithmic

transformation.

TABLE 5
Skewness, kurtosis and values of the S-W p-values for the municipal- and household-level error terms of the working models for EBP under the different transformations

	Household-level residuals			Municipal-level residuals		
Transformation	Skewness	Kurtosis	p-value	Skewness	Kurtosis	p-value
No	7.981	110.697	0.000	1.298	5.596	0.002
Log	−1.480	6.653	0.000	−0.576	2.336	0.025
Log-Shift	−0.346	3.895	0.000	−0.057	1.969	0.226
Box-Cox	−0.118	5.311	0.000	−0.023	2.181	0.484
Dual	−0.024	5.809	0.000	−0.005	2.242	0.627

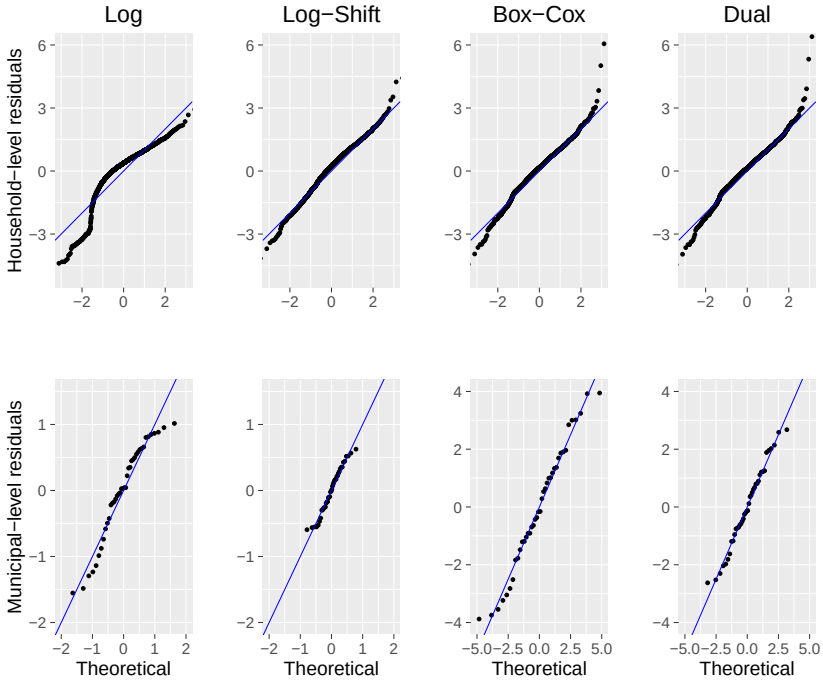

FIG 4. *Q-Q-plots for the Pearson household-level (upper panels) and municipal-level (lower panels) residuals of the working model for EBP under the different transformations*

6.2. *Deprivation and inequality indicators for municipalities in Guerrero.* Based on the analysis in Section 6.1, estimates for the deprivation and inequality indicators presented in Section 2 are calculated by using the EBP

method under the three data-driven transformations and the fixed logarithmic transformation. MSE estimation is implemented with the wild bootstrap we introduced in Section 5 with $B = 500$ bootstrap replications.

Table 6 shows summaries over municipalities of point estimates and root MSEs (RMSEs) under the different transformations. We observe that the estimates based on the EBP with data-driven transformations are more efficient (in terms of RMSE) than the corresponding estimates based on a fixed logarithmic transformation. It appears that transformations suggested by the model checking and residual diagnostics lead to more efficient estimates. The effect is especially pronounced for indicators that rely on the tail of the distribution like the QSR. Note that we report in the application a modified QSR calculated as the ratio of total income received by the 40% of the households with the highest income divided by the total income received by the 40% of the households with the lowest income. Furthermore, the use of data-driven transformations also has an effect on the point estimates of the indicators. Especially for the HCR and PGAP, the three data-driven transformations result in very similar results, which are different to the EBP estimates under the model that uses the logarithmic transformation.

TABLE 6
Summaries of point estimates and corresponding RMSEs over municipalities in Guerrero

Point Estimation	HCR		PGAP		QSR	
Transformation	Mean	Median	Mean	Median	Mean	Median
Log	0.64	0.66	0.46	0.47	56.03	54.64
Log-Shift	0.56	0.59	0.35	0.36	18.06	15.83
Box-Cox	0.55	0.57	0.37	0.38	23.53	22.71
Dual	0.54	0.57	0.37	0.38	27.79	25.11
RMSE	HCR		PGAP		QSR	
Transformation	Mean	Median	Mean	Median	Mean	Median
Log	0.12	0.12	0.11	0.13	90.96	86.23
Log-Shift	0.10	0.11	0.09	0.09	8.73	5.92
Box-Cox	0.10	0.10	0.09	0.09	7.03	6.11
Dual	0.09	0.10	0.09	0.09	7.71	6.55

Having assessed the estimates from a statistical perspective, we investigate the results in the context of spatial distribution of poverty and inequality in the state Guerrero. Figure 5 presents the point estimates of HCR, PGAP and QSR at municipal-level. As the point estimates based on the three data-driven transformations are almost identical, we only show here the results for the EBP with log-shift transformation. We observe clearly regional differences between the municipalities. Having a closer look to the

coastal area in the south-west of Guerrero, where the largest city of Guerrero, Acapulco, is located, we observe lower levels of poverty (HCR and PGAP) and inequality (QSR) compared to other parts of the state. The coastline to the Pacific Ocean is wealthier due to several tourist destinations like Acapulco, Ixtapa and Zihuatanejo. In contrast, there is also a clear hotspot in the eastern part of the state Guerrero (e.g. municipalities: Metlatnoc, Malinaltepec and Atlixtac) with high poverty and inequality rates. These municipalities are the home of ethnic groups indigenous and most of them live in isolated mountain areas.

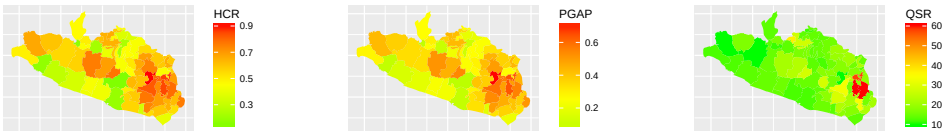

FIG 5. *Maps of the HCR, PGAP and QSR in Guerrero for the EBP method under the log-shift transformation at municipal-level*

7. Model-based simulation study. In this section, we present results from a model-based simulation study in order to evaluate the performance of the EBP method under data-driven transformations presented in Section 4. The objective of the simulation study is fourfold: Firstly, we analyze the behaviour of the data-driven transformation parameter under four different distributions in Section 7.1. Secondly, we investigate the ability of the proposed EBP method under data-driven transformations to account for different shapes of distributions in Section 7.2, and hence provide more precise small area estimates than the EBP with a fixed logarithmic or without a transformation. Thirdly, we discuss the performance of the proposed MSE estimators introduced in Section 5. Finally, the sensitivity of the proposed methodology in relation to the estimation method for the data-driven transformation parameter is assessed in Section 7.3.

We generate finite populations U of size $N = 10000$, partitioned into $D = 50$ areas $U_1, U_2, \ldots, U_D$ of sizes $N_i = 200$. The samples are selected by a stratified random sampling with strata defined by the 50 small areas. This leads to a sample size of $n = \sum_{i=1}^{D} n_i = 921$ whereby the area-specific sample sizes n_i vary between 8 and 29. Four scenarios, denoted by *Normal*, *Log-scale*, *Pareto* and *GB2*, are considered. Details about the data generating mechanisms of the different scenarios are provided in Table 7. Under scenario *Normal*, data are generated by using Normal distributions for the random effects and error terms - using untransformed data should be ap-

propriate. In contrast, the second scenario *Log-scale* generates data under a log-normal distribution such that a fixed logarithmic transformation is suitable. Scenarios *Pareto* and *GB2* are settings where the data are right-skewed, like in the case of income distributions. Each setting was repeated independently $M = 500$ times. We focus on the three data-driven transformations presented in Section 4.2, namely log-shift, Box-Cox and dual power transformations, compared to a fixed logarithmic transformation and without a transformation.

TABLE 7
Model-based simulation settings for the analysis of the MSE

Scenario	Model	x_{ij}	z_{ij}	μ_i	u_i	e_{ij}
Normal	$4500 - 400x_{ij} + u_i + e_{ij}$	$N(\mu_i, 3)$	-	$U[-3, 3]$	$N\left(0, 500^2\right)$	$N(0, 1000^2)$
Log-scale	$\exp(10 - x_{ij} - 0.5z_{ij} + u_i + e_{ij})$	$N(\mu_i, 2)$	$N(0, 1)$	$U[2, 3]$	$N\left(0, 0.4^2\right)$	$N(0, 0.8^2)$
Pareto	$12000 - 400x_{ij} + u_i + e_{ij} - \bar{e}$	$N(\mu_i, 7.5)$	-	$U[-3, 3]$	$N\left(0, 500^2\right)$	$\sqrt{2}\text{Pareto}(3, 2000^2)$
GB2	$8000 - 400x_{ij} + u_i + e_{ij} - \bar{e}$	$N(\mu_i, 5)$	-	$U[-1, 1]$	$N\left(0, 500^2\right)$	$\text{GB2}(2.5, 1700, 18, 1.46)$

7.1. *Behaviour of the data-driven transformation parameters.* Before we assess the performance of the EBP under different data-driven transformations in Section 7.2, we have a first look to the behaviour of the estimated transformation parameters in Figure 6. In particular, the figure shows the box-plots of the estimated transformation parameters λ for log-shift, Box-Cox and dual power transformations (over $M = 500$ replications) for the four settings presented in Table 7. The data-driven transformation parameters are estimated by REML introduced in Section 4.2. To start with, in the *Normal* setting, the parameters of the Box-Cox and dual power transformations tend to one indicating that no transformation is needed. In the *Log-scale* scenario, the data was generated in such a way that normality may be achieved by applying the logarithmic transformation. As expected, the log-shift transformation parameters tend to zero, for which the log-shift transformation is equivalent to the logarithmic transformation. The same behaviour can be observed for the Box-Cox and dual power transformations, as the estimated transformation parameters are close to zero, corresponding to using the logarithmic transformation. For the other two scenarios (*Pareto* and *GB2*), the data-driven parameters lie between 0.25 and 0.5, so neither using a logarithmic transformation or no transformation seems to be appropriate. Note, the log-shift transformation parameter has only a natural interpretation for $\lambda = 0$ as the shift-parameter depends on the scale of the data.

Overall, the results indicate that the data-driven transformations behave as expected in the four scenarios and adapt to the different shape of the distributions.

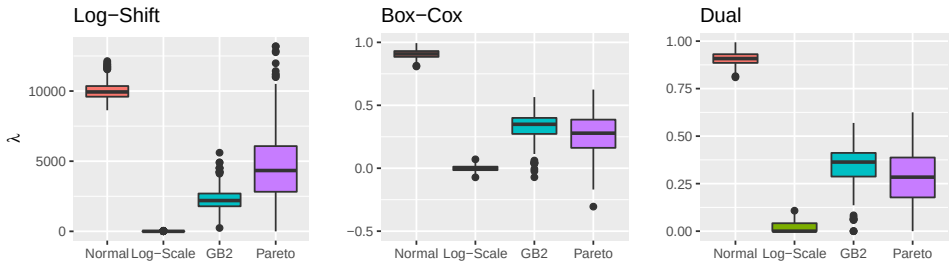

FIG 6. *Estimated transformation parameters for the log-shift, Box-Cox and dual power transformations under the different settings.*

7.2. *Performance of the EBP under data-driven transformations.* In this section, we evaluate the performance of the proposed EBP method under data-driven transformations compared to a fixed logarithmic and without a transformation. Afterwards, we assess the performance of the introduced MSE estimators.

Five estimators of small area population indicators are evaluated. These are the EBP without transformation introduced in Section 2 and the EBP with a fixed logarithmic transformation. Furthermore, we analyze the EBP with three data-driven transformations (log-shift, Box-Cox and dual power transformations). For estimating the EBPs and the corresponding MSE estimates, the parameters L and B are set to 100 and 500, respectively. The choice is justifiable, as Molina and Rao [1] suggest that a choice around 50 gives fairly accurate results. The following quality measures, over Monte-Carlo replications M, are used to assess the performance of a small area estimator in area i:

$$\text{RMSE}\left(\hat{I}^{\text{method}}\right) = \left[\frac{1}{M}\sum_{m=1}^{M}\left(\hat{I}_i^{\text{method}(m)} - I_i^{(m)}\right)^2\right]^{1/2},$$

$$\text{Bias}\left(\hat{I}^{\text{method}}\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{I}_i^{\text{method}(m)} - I_i^{(m)}\right),$$

where $\hat{I}_i^{\text{method}}$ denotes an estimated indicator in area i based on any of the five methods discussed above and I_i denotes the corresponding true value in area i. To be precise, we evaluate three different indicators I_i (HCR, PGAP and QSR) which tackle different parts of the distribution. The indicators HCR and PGAP depend on a poverty line t which is equal to 0.6 times the median of the target variable. In contrast, the QSR depends on the lower and upper 20% of the estimated distribution and is sensitive to the tails.

TABLE 8
Summaries of estimated RMSEs and Bias over the model-based settings

	Indicator	HCR		PGAP		QSR	
	Estimator	Median	Mean	Median	Mean	Median	Mean
Normal							
RMSE	No	0.0338	0.0357	0.0136	0.0154	0.3259	1.2765
	Log-Shift	0.0344	0.0363	0.0155	0.0175	0.3898	0.6710
	Box-Cox	0.0343	0.0358	0.0134	0.0156	0.3348	1.1178
	Dual	0.0343	0.0358	0.0134	0.0156	0.3346	0.5797
BIAS	No	0.0000	0.0007	0.0002	0.0009	0.0049	0.0899
	Log-Shift	0.0029	0.0039	-0.0067	-0.0076	-0.1000	-0.2190
	Box-Cox	0.0016	0.0027	-0.0021	-0.0025	-0.0396	-0.0807
	Dual	0.0016	0.0027	-0.0021	-0.0024	-0.0458	-0.1193
Log-Scale							
RMSE	Log	0.0583	0.0605	0.0358	0.0367	4.9100	4.8969
	Log-Shift	0.0583	0.0605	0.0358	0.0367	4.9024	4.8985
	Box-Cox	0.0581	0.0604	0.0358	0.0367	4.9731	4.9717
	Dual	0.0584	0.0605	0.0359	0.0367	4.9025	4.9093
BIAS	Log	-0.0011	-0.0009	-0.0007	-0.0003	0.0394	0.1143
	Log-Shift	-0.0020	-0.0017	-0.0011	-0.0007	-0.0873	-0.0072
	Box-Cox	-0.0009	-0.0006	-0.0008	-0.0004	0.1499	0.2106
	Dual	-0.0024	-0.0021	-0.0009	-0.0005	-0.1610	-0.0992
GB2							
RMSE	No	0.0650	0.0656	0.0552	0.0552	17.7364	32.0686
	Log	0.0912	0.0908	0.0272	0.0270	1.8979	1.9002
	Log-Shift	0.0418	0.0415	0.0127	0.0132	0.4286	0.4411
	Box-Cox	0.0471	0.0469	0.0136	0.0139	0.4708	0.4753
	Dual	0.0472	0.0470	0.0137	0.0140	0.4715	0.4760
BIAS	No	0.0471	0.0477	0.0481	0.0479	1.8355	2.0825
	Log	0.0746	0.0747	0.0169	0.0169	1.4718	1.4692
	Log-Shift	0.0176	0.0179	-0.0008	-0.0013	0.0546	0.0523
	Box-Cox	0.0274	0.0274	0.0035	0.0031	0.1780	0.1721
	Dual	0.0275	0.0274	0.0037	0.0034	0.1800	0.1747
Pareto							
RMSE	No	0.0448	0.0444	0.0622	0.0613	1.6814	3.6057
	Log	0.0304	0.0306	0.0082	0.0084	0.3887	0.3994
	Log-Shift	0.0185	0.0196	0.0060	0.0063	0.1661	0.1779
	Box-Cox	0.0192	0.0202	0.0059	0.0062	0.1786	0.1901
	Dual	0.0192	0.0203	0.0059	0.0062	0.1782	0.1902
BIAS	No	0.0277	0.0287	0.0166	0.0160	0.3173	0.3132
	Log	0.0086	0.0081	-0.0030	-0.0037	0.2068	0.2034
	Log-Shift	0.0003	-0.0001	-0.0034	-0.0041	0.0305	0.0300
	Box-Cox	0.0030	0.0026	-0.0031	-0.0037	0.0525	0.0530
	Dual	0.0030	0.0027	-0.0031	-0.0037	0.0522	0.0530

TABLE 9
Performance of MSE estimators in model-based simulations: EBP with Box-Cox transformation

	Indicator	HCR		PGAP		QSR	
	Estimator	Median	Mean	Median	Mean	Median	Mean
Normal							
rel. RMSE[%]	Parametric	8.30	9.22	9.15	9.47	15.25	21.23
	Wild	14.57	14.77	14.21	14.61	17.46	20.93
rel. Bias[%]	Parametric	6.64	7.27	-1.17	-0.12	-7.72	-12.61
	Wild	8.05	8.04	2.17	3.23	-1.01	-1.46
Log-Scale							
rel. RMSE[%]	Parametric	11.14	12.00	19.19	19.57	19.10	19.75
	Wild	16.82	17.00	22.70	22.95	25.34	25.62
rel. Bias[%]	Parametric	6.10	6.29	5.70	6.36	7.91	7.92
	Wild	7.69	7.82	7.34	7.39	6.58	6.78
GB2							
rel. RMSE[%]	Parametric	21.71	21.86	20.89	20.57	43.75	43.58
	Wild	19.01	19.39	14.76	15.12	26.21	27.23
rel. Bias[%]	Parametric	-20.04	-19.74	-16.88	-15.92	-42.90	-42.74
	Wild	-14.59	-14.64	-5.45	-5.75	-21.72	-22.53
Pareto							
rel. RMSE[%]	Parametric	11.31	12.60	35.60	34.78	50.04	51.63
	Wild	26.18	28.44	23.58	26.04	28.60	33.40
rel. Bias[%]	Parametric	2.43	3.38	-33.82	-31.16	-49.51	-51.06
	Wild	19.21	21.37	-8.28	-3.28	-23.02	-26.79

Table 8 presents the results split by the four scenarios. The table presents median and mean values of RMSE and bias over small area. Under the *Normal* scenario the EBP without transformation is the gold standard. However, the EBP with data-driven transformations (log-shift, Box-Cox and dual power) perform similar in terms of RMSE and bias. It can be observed that all estimators are almost unbiased in the *Normal* scenario. The same picture holds in the *Log-scale* scenario where the EBP with a logarithmic transformation is the gold standard. Again, it seems the EBP with data-driven transformations perform more or less on the same level in terms of RMSE and bias for all three indicators. The results confirm our expectations that the EBP with data-driven transformations adapt to the shape of the underlying distributions in the *Normal* and *Log-scale* settings and perform similarly compared to the EBP with optimal transformation. Under *GB2* and *Pareto* scenarios we notice that the EBP with fixed transformations (either a logarithmic or without a transformation) is inferior to the EBP

with data-driven transformations in terms of RMSE and Bias for all three indicators. The differences are especially pronounced for the QSR which is very sensitive to the tails of the distribution, whereas the HCR and PGAP only depend on the lower quantiles. Furthermore, the estimates based on data-driven transformations are almost unbiased or reveal only a small bias for the three indicators. A closer look at the data-driven transformations might indicate that EBP with a log-shift transformation perform somewhat better compared to the EBP with Box-Cox and dual power transformations in these particular settings. Overall, it seems that the proposed EBP method with data-driven transformations account for different shapes of distributions, and hence provide more precise small area estimates than the EBP with a fixed logarithmic or without a transformation in the four particular settings.

We now turn to the performance of the different MSE estimators. We denote by *parametric* and *wild* the proposed parametric bootstrap and proposed semi-parametric wild bootstrap respectively. The aim of this part is twofold: Firstly, we assess the performance of the two proposed MSE estimators introduced in Section 5. Secondly, we investigate the potential feature of the wild bootstrap to protect against departures from the assumptions of the unit-level error term. Starting with the first aim, Table 9 reports the results for the two MSE estimators and presents the mean and median values of relative RMSE and relative Bias for the EBP with Box-Cox transformation. We treat the empirical MSE (over Monte-Carlo replications) as the true MSE. The corresponding results for the EBP with a log-shift transformation and dual power transformation are available on request from the authors. We note that, on average, the proposed *parametric* and *wild* bootstrap approaches for the EBP with a Box-Cox transformation are almost unbiased for the HCR and PGAP in the *Normal* and *Log-scale* settings. However, the *parametric* bootstrap schemes shows some underestimation (in terms of rel. Bias) for QSR. Overall, it seems that both bootstrap approaches lead to reasonable results, provided the population model is correct. Under the *GB2* and *Pareto* settings, both bootstraps show a negative bias, especially for the QSR. Nevertheless, the *wild* bootstrap provides reasonable results for HCR and PGAP and reduces the underestimation for QSR. The results indicate that even small departures from the model assumptions can have an adverse effect on MSE estimation based on the *parametric* bootstrap of non-linear indicators, computation of which depends on the entire target distribution like QSR. In contrast, the *wild* bootstrap protects somehow against small departures from the model assumptions of the unit-level error term and provides more conservative results than the *parametric* bootstrap scheme in the

simulation study presented here.

7.3. *Impact of alternative estimation methods for* λ. In Section 4.3 we proposed non-parametric alternatives to the REML method in order to obtain data-driven transformation parameters. Here, we briefly discuss the impact of five estimation methods. These are the REML approach introduced in Section 4.2, the minimization of the skewness (*Skew*) and the pooled skewness (*poolSkew*), and the distance-based criteria Kolomogorov-Smirnon (*KS*) and the Cramér-von Mises (*CvM*) defined in Section 4.3. In our particular simulation scenarios, introduced in Table 7, we observed that the resulting point and MSE estimates are only marginally influenced by the estimation method of the transformation parameter. The five methods estimate parameters close to the theoretically correct ones, in the scenarios those are known. For instance, in the *Log-scale* scenario, the estimated transformation parameters under the different estimation methods are shown in Figure 7. We observe that the five methods result in similar estimates for the transformation parameter λ, but the REML method, discussed in detail before, tends to have a smaller variability. Table 10 shows the mean and median values of the estimated transformation parameters. From this, it becomes clear that all estimation methods result on average into parameters that are very close to the theoretically correct ones in this particular scenario.

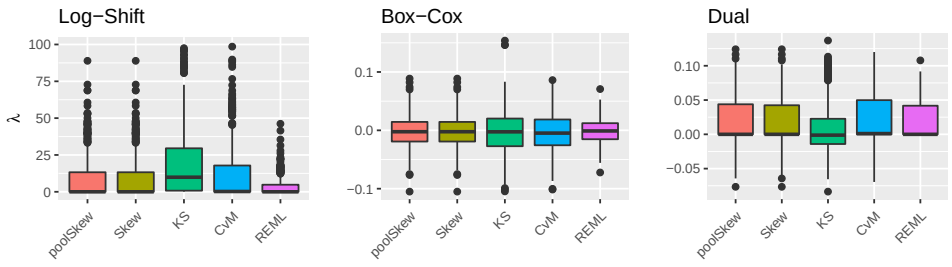

FIG 7. *Box-plots of estimated transformation parameters for the log-scale scenario using different estimation methods*

8. Conclusions and future research directions. In this paper, we proposed data-driven transformations for small area estimation. In particular, we introduced the EBP approach with data-driven transformations in general and proposed a likelihood-based approach for estimating the adaptive transformation parameter. The scaled transformations are conditional on the Jacobian, such that standard software procedures for estimating the

TABLE 10
Mean and median of estimated transformation parameters under the log-scale scenario using different estimation methods

	Log-Shift		Box-Cox		Dual	
	Mean	Median	Mean	Median	Mean	Median
poolSkew	9.381	0.000	-0.002	-0.002	0.016	0.000
Skew	9.381	0.000	-0.002	-0.002	0.015	0.000
KS	23.906	10.816	-0.003	-0.003	0.009	-0.001
CvM	11.954	0.211	-0.004	-0.005	0.025	0.001
REML	3.349	0.000	-0.002	-0.001	0.021	0.000

optimal transformation parameter can be used. Although the theory is introduced for data-driven transformations in general, we additionally discuss three subcases (log-shift, Box-Cox and dual power transformations). As the likelihood-based approaches are based on parametric assumptions, we also propose non-parametric alternatives for estimating the adaptive transformation parameter. Model-based simulations demonstrate the ability of the proposed EBP method with data-driven transformations to account for different shapes of distributions and, hence, provide more efficient results compared to using a fixed logarithmic transformation. Although the paper focuses on the EBP as a specific small area estimator, the proposed data-driven transformations are applicable to other small area estimators, for example the ELL [2].

However, even if the optimal data-driven transformation of the data has been found, there may still be departures from model assumptions. Such departures can affect the quality of the small area estimates and can impact the quality of the precision estimates - in terms of MSE - based on the use of a parametric bootstrap. Therefore, we also proposed a semi-parametric wild bootstrap that (a) protects against departures from model assumptions in particular, those of the unit-level error term and (b) captures the additional uncertainty coming from the estimation of the data-driven transformation parameter. Finally, we demonstrated the need for data-driven transformations in an application based on data from the state Guerrero in Mexico by estimating poverty and inequality indicators for 81 municipalities.

Further research should investigate additional transformation families, especially multi-parameter families. As different indicators might me more sensitive to different parts of the distributions (center or tails), this may allow for a better control of higher moments, like the kurtosis, and could lead to potentially more efficient results. Additionally, further research could compare the parametric and non-parametric methods for estimating the data-driven

transformation parameter in more detail. As likelihood-based approaches might be influenced by outliers in data, it would be interesting to investigate the robust estimation methods proposed by [35] and to study the influence of outliers and their effect after deletion in the context of transformations following [36, 37]. Finally, we discussed one possible approach to accounting for departures from normality in the EBP method proposed by Molina and Rao [1]. Alternative approaches formulate the EBP under more flexible parametric assumptions [6, 7] or use methodology that minimises the use of parametric assumptions [8]. A detailed comparison of the different approaches under realistic settings is an open research problem.

APPENDIX A: LIKELIHOOD DERIVATION OF THE TRANSFORMATIONS

We derive the Jacobian and the corresponding scaled data-driven transformations presented in Table 3 for the log-shift, Box-Cox, and dual power transformations, as outlined below.

A.1. Log-Shift transformation. Let $J(\lambda, \boldsymbol{y})$ be the Jacobian of the log-shift transformation presented in Section 4 from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^*(\lambda)$, defined as:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{D}\prod_{j=1}^{n}\left|\frac{dy_{ij}^*(\lambda)}{dy_{ij}}\right| \\ &= \prod_{i=1}^{D}\prod_{j=1}^{n}(y_{ij}+\lambda)^{-1}. \end{aligned}$$

Therefore, the log-likelihood function given in 4.2 is re-defined as:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}] \\ &\quad \times n(-1)\log\left(\prod_{i=1}^{D}\prod_{j=1}^{n_i} y_{ij+\lambda}\right)^{\frac{1}{n}}. \end{aligned}$$

Taking the definition of the geometric mean of a variable:

$$\bar{y} = \left(\prod_{i=1}^{D} \prod_{j=1}^{n_i} y_{ij} \right)^{\frac{1}{n}}.$$

In case of using the log-shift transformation, $\bar{y}$ is denoted by:

$$\bar{y} = \left(\prod_{i=1}^{D} \prod_{j=1}^{n_i} y_{ij} + \lambda \right)^{\frac{1}{n}}.$$

The log-likelihood function presented is re-written as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D} \boldsymbol{X}_i^T \boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T \boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}] \\ &\quad \times n(-1)\log(\bar{y}). \end{aligned}$$

In order to obtain the scaled log-shift transformation, $z_{ij}^*(\lambda)$, the denominator of the term $\frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{1/n}}$ is given as:

$$\begin{aligned} 1/J(\lambda, \boldsymbol{y})^{\frac{1}{n}} &= \left[\left\{ \left(\prod_{i=1}^{D} \prod_{j=1}^{n_i} y_{ij} \right)^{\frac{1}{n}} \right\}^n \right]^{\frac{1}{n}} \\ &= \bar{y}. \end{aligned}$$

Therefore, the scaled log-shift transformation is defined as follows:

$$z_{ij}^*(\lambda) = \bar{y}\ \log(y_{ij} + \lambda)$$

for $y_{ij} > 0$.

A.2. Box-Cox transformation. Let $J(\lambda, \boldsymbol{y})$ be the Jacobian of the Box-Cox transformation presented in Section 4 from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^*(\lambda)$, defined as:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{D}\prod_{j=1}^{n}\left|\frac{dy_{ij}^*(\lambda)}{dy_{ij}}\right| \\ &= \prod_{i=1}^{D}\prod_{j=1}^{n} y_{ij}^{\lambda-1}. \end{aligned}$$

Therefore, the log-likelihood function given in 4.2 is re-defined as:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}] \\ &\quad \times n(\lambda-1)\log\left(\prod_{i=1}^{D}\prod_{j=1}^{n_i} y_{ij}\right)^{\frac{1}{n}}. \end{aligned}$$

Taking the definition of the geometric mean of a variable:

$$\bar{y} = \left(\prod_{i=1}^{D}\prod_{j=1}^{n_i} y_{ij}\right)^{\frac{1}{n}}.$$

The log-likelihood function presented is re-written as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}] \\ &\quad \times n(\lambda-1)\log(\bar{y}). \end{aligned}$$

In order to obtain the scaled transformation of the Box-Cox family, $z_{ij}^{*}(\lambda)$, the denominator of the term $\frac{y_{ij}^{*}(\lambda)}{J(\lambda,\boldsymbol{y})^{1/n}}$ is given as:

$$\begin{aligned} 1/J(\lambda,\boldsymbol{y})^{\frac{1}{n}} &= \left[\left\{\left(\prod_{i=1}^{D}\prod_{j=1}^{n_i} y_{ij}\right)^{\frac{1}{n}}\right\}^{-n(\lambda-1)}\right]^{\frac{1}{n}} \\ &= \bar{y}^{-(\lambda-1)}. \end{aligned}$$

Therefore, the scaled Box-Cox transformation is defined as follows:

$$z_{ij}^{*}(\lambda) = \begin{cases} \frac{y_{ij}^{\lambda}-1}{\bar{y}^{\lambda-1}\lambda} & \text{if } \lambda \neq 0, \\ \bar{y}\log(y_{ij}) & \text{if } \lambda = 0. \end{cases}$$

for $y_{ij} > 0$. Including the shift parameter s, the scaled Box-Cox transformation would be written as:

$$z_{ij}^{*}(\lambda) = \begin{cases} \frac{(y_{ij}+s)^{\lambda}-1}{\bar{y}^{\lambda-1}\lambda}, & \lambda \neq 0, \\ \bar{y}\log(y_{ij}+s), & \lambda = 0, \end{cases}$$

whereby the geometrical mean of the scaled Box-Cox transformation is defined as:

$$\bar{y} = \prod_{i=1}^{D}\prod_{j=1}^{n}(y_{ij}+s)^{\frac{1}{n}}$$

A.3. Dual power transformation. Let $J(\lambda,\boldsymbol{y})$ be the Jacobian of the dual power transformation from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^{*}(\lambda)$, defined as:

$$\begin{aligned} J(\lambda,\boldsymbol{y}) &= \prod_{i=1}^{D}\prod_{j=1}^{n}\left|\frac{dy_{ij}^{*}(\lambda)}{dy_{ij}}\right| \\ &= \frac{\prod_{i=1}^{D}\left(\prod_{j=1}^{n} y_{ij}^{\lambda-1} + y_{ij}^{-\lambda-1}\right)}{2}, \end{aligned}$$

In case of using the dual transformation, $\bar{y}$ is denoted by:

$$\bar{y} = \begin{cases} \left(\prod_{i=1}^{D}\prod_{j=1}^{n} y_{ij}^{\lambda-1} + y_{ij}^{-\lambda-1}\right)^{\frac{1}{n}} & \text{if } \lambda > 0, \\ \left(\prod_{i=1}^{D}\prod_{j=1}^{n_i} y_{ij}\right)^{\frac{1}{n}} & \text{if } \lambda = 0. \end{cases}$$

In case $\lambda = 0$, the derivation below is equivalent to the second case of the Box-Cox ($\lambda = 0$). In case $\lambda > 0$ the log-likelihood function can be defined as:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D} \boldsymbol{X}_i^T \boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T \boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}] \\ &\times \frac{n}{2}\log\left(\prod_{i=1}^{D}\prod_{j=1}^{n} y_{ij}^{\lambda-1} + y_{ij}^{-\lambda-1}\right)^{\frac{1}{n}}. \end{aligned}$$

Taking the term in the denominator of the dual transformation, it holds:

$$\begin{aligned} 1/J(\lambda, \boldsymbol{y})^{1/n} &= \left[\left[\left(\frac{\prod_{i=1}^{D}\prod_{j=1}^{n_i} y_{ij}^{\lambda-1} + y_{ij}^{-\lambda-1}}{2}\right)^{1/n}\right]^{-n}\right]^{\frac{1}{n}} \\ &= 2\bar{y}^{-1} \end{aligned}$$

Therefore, incorporating the fixed parameter s, the scaled dual power transformation is defined as follows:

$$z_{ij}^*(\lambda) = \begin{cases} 2\bar{y}^{-1}\frac{(y_{ij}+s)^{\lambda} - (y_{ij}+s)^{-\lambda}}{2\lambda} & \text{if } \lambda > 0; \\ \bar{y}^{-1}\log(y_{ij}+s) & \text{if } \lambda = 0. \end{cases}$$

REFERENCES

[1] I. Molina and J. N. K. Rao. Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38(3):369–385, 2010.

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

[3] E. Fabrizi and C. Trivisano. Small area estimation of the gini concentration coefficient. *Computational Statistics & Data Analysis*, 99:223–234, 2016.

[4] T. Schmid, F. Bruckschen, N. Salvati, and T. Zbiranski. Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in senegal. *Journal of the Royal Statistical Society: Series A*, 180(4):1163–1190, 2017.

[5] I. Molina and J. N. K. Rao. A review of poverty mapping procedures. *Proceedings 59th ISI World Statistics Congress, Hong Kong (Session IPS080)*, 2013.

[6] M. Graf, J. Marin, and I. Molina. Estimation of poverty indicators in small areas under skewed distributions. Working paper, 2014. Accessed: 2016-02-16.
[7] M. S. Diallo and J. N. K. Rao. Small area estimation of complex parameters under unit-level models with skew-normal errors. JSM 2014, Survey Research Methods Section, 2014.
[8] B. Weidenhammer, N. Tzavidis, T. Schmid, and N. Salvati. Domain prediction for counts using microsimulation via quantiles. In *Small Area Estimation 2014 Conference*, 2014.
[9] M. F. Marino, N. Tzavidis, and M. Alfo. Mixed hidden markov quantile regression models for longitudinal data with possibly incomplete sequences. *Statistical Methods in Medical Research*, page forthcoming, 2016.
[10] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, 26(2):211–252, 1964.
[11] B. F. J. Manly. Exponential data transformations. *Journal of the Royal Statistical Society: Series D*, 25(1):37–42, 1976.
[12] J. A. John and N. R. Draper. An alternative family of transformations. *Journal of the Royal Statistical Society: Series C*, 29(2):190–197, 1980.
[13] P. J. Bickel and K. A. Doksum. An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296 – 311, 1981.
[14] I.-K. Yeo and R. A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.
[15] H. Cramér. On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928:13–74, 1928.
[16] I. M. Chakravarti and R. G. Laha. Handbook of methods of applied statistics. In *Handbook of methods of applied statistics*. John Wiley & Sons, 1967.
[17] R. J. Carroll and D. Ruppert. Diagnostics and robust estimation when transforming the regression model and the response. *Technometrics*, 29(3):287–299, 1987.
[18] G. E. Battese, R. M. Harter, and W. A. Fuller. An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36, 1988.
[19] J. Foster, J. Greer, and E. Thorbecke. A class of decomposable poverty measures. *Econometrica*, 52(3):761–766, 1984.
[20] Eurostat. Common cross-sectional eu indicators based on eu-silc; the gender pay gap. *Unit D-2: Living conditions and social protection, Directorate D: Single Market, Employment and Social statistics, Eurostat, Luxembourg*, 2004.
[21] C. Gini. Variabilità e mutabilità : Contributo allo studio e delle distribuzioni e relazioni statistiche. *Studi Economico-Giuridici della R, Universitá di Cagliari*, 1912.
[22] S. Nakagawa and H. Schielzeth. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2): 133–142, 2013.
[23] H. Thoni. Transformation of variables used in the analysis of experimental and observational data: A review. *Journal of the American Statistical Association*, 64 (327):1099, 1969.
[24] G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*, volume 1. Springer Series in Statistics, 2000.
[25] D. A. Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2):383–385, 1974.
[26] Q. Feng, J. Hannig, and J. S. Marron. A note on automatic data transformation. *Stat*, 2016.
[27] Z. Yang. A modified family of power transformations. *Economics Letters*, 92:14–19,

2006.

[28] D. V. Hinkley. On power transformations to symmetry. *Biometrika*, 62:101–111, 1975.

[29] R. Sakia. The box-cox transformation technique: A review. *Journal of the Royal Statistical Society: Series D*, 41(2):169–178, 1992.

[30] P. Royston and P. C. Lambert. *Flexible parametric survival analysis using Stata: Beyond the Cox model*. StataCorp LP, 2011.

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

[32] H.-T. Thai, F. Mentré, N. H. Holford, C. Veyrat-Follet, and E. Comets. A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharmaceutical statistics*, 12(3):129–140, 2013.

[33] E. Flachaire. Bootstrapping heteroskedastic regression models: Wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis*, 49(2):361–376, 2005. 2nd {CSDA} Special Issue on Computational Econometrics.

[34] X. Feng, X. He, and J. Hu. Wild bootstrap for quantile regression. *Biometrika*, 98 (4):995, 2011.

[35] R. J. Carroll. A robust method for testing transformations to achieve approximate normality. *Journal of the Royal Statistical Society: Series B*, 42(1):71–78, 1980.

[36] R. Cook and P. Wang. Transformations and influential cases in regression. *Technometrics*, 25(4):337–343, 1983.

[37] A. C. Atkinson. Diagnostic tests for transformations. *Technometrics*, 28(1):29–37, 1986.

Rojas-Perilla Natalia
Institute of Statistics and Econometrics
Freie Universität Berlin
Germany
E-mail: natalia.rojas@fu-berlin.de

Sören Pannier
Institute of Statistics and Econometrics
Freie Universität Berlin
Germany
E-mail: soeren.pannier@fu-berlin.de

Timo Schmid
Institute of Statistics and Econometrics
Freie Universität Berlin
Germany
E-mail: timo.schmid@fu-berlin.de

Nikos Tzavidis
Southampton Statistical Sciences Research Institute
University of Southampton
UK
E-mail: n.tzavidis@soton.ac.uk

Diskussionsbeiträge - Fachbereich Wirtschaftswissenschaft - Freie Universität Berlin
Discussion Paper - School of Business and Economics - Freie Universität Berlin

2017 erschienen:

2017/1 ARONSSON, Thomas und Ronnie SCHÖB
Habit Formation and the Pareto-Efficient Provision of Public Goods
Economics

2017/2 VOGT, Charlotte; Martin GERSCH und Cordelia GERTZ
Governance in integrierten, IT-unterstützten Versorgungskonzepten im Gesundheitswesen : eine Analyse aktueller sowie zukünftig möglicher Governancestrukturen und -mechanismen
Wirtschaftsinformatik

2017/3 VOGT, Charlotte; Martin GERSCH und Hanni KOCH
Geschäftsmodelle und Wertschöpfungsarchitekturen intersektoraler, IT-unterstützter Versorgungskonzepte im Gesundheitswesen
Wirtschaftsinformatik

2017/4 DOMBI, Akos und Theocharis GRIGORIADIS
Ancestry, Diversity & Finance : Evidence from Transition Economies
Economics

2017/5 SCHREIBER, Sven
Weather Adjustment of Economic Output
Economics

2017/6 NACHTIGALL, Daniel
Prices versus Quantities: The Impact of Fracking on the Choice of Climate Policy Instruments in the Presence of OPEC
Economics

2017/7 STOCKHAUSEN, Maximilian
The Distribution of Economic Resources to Children in Germany
Economics

2017/8 HETSCHKO, Clemens; Louisa von REUMONT und Ronnie SCHÖB
Embedding as a Pitfall for Survey-Based Welfare Indicators: Evidence from an Experiment
Economics

2017/9 GAENTZSCH, Anja
Do Conditional Cash Transfers (CCT) Raise Educational Attainment? A Case Study of Juntos in Peru
Economics

2017/10 BACH, Stefan; Martin BEZNOSKA und Viktor STEINER
An Integrated Micro Data Base for Tax Analysis in Germany
Economics

2017/11 NEUGEBAUER, Martin und Felix WEISS
Does a Bachelor's Degree pay off? Labor Market Outcomes of Academic versus Vocational Education after Bologna
Economics

2017/12 HACHULA, Michael und Dieter NAUTZ
The Dynamic Impact of Macroeconomic News on Long-Term Inflation Expectations
Economics

2017/13 CORNEO, Giacomo
Ein Staatsfonds, der eine soziale Dividende finanziert
Economics

2017/14 GERSCH, Martin; Cordelia GERTZ und Charlotte VOGT
Leistungsangebote in integrierten, IT-unterstützten Versorgungskonzepten: eine Konzeption (re-) konfigurierbarer Servicemodule im Gesundheitswesen
Wirtschaftsinformatik

2017/15 KREUTZMANN, Ann-Kristin; Sören PANNIER; Natalia ROJAS-PERILLA; Timo SCHMID; Matthias TEMPL und Nikos TZAVIDIS
The R Package emdi for Estimating and Mapping
Regionally Disaggregated Indicators
Economics

2017/16 VOGT, Charlotte; Cordelia GERTZ und Martin GERSCH
Ökonomische Evaluation eines integrierten, IT-unterstützten Versorgungskonzepts im Gesundheitswesen: eine ökonomische Analyse von E-Health-unterstützten Versorgungsprozessen
Wirtschaftsinformatik

2017/17 GASTEIGER, Emanuel und Klaus PRETTNER
A Note on Automation, Stagnation, and the Implications of a Robot Tax
Economics

2017/18 HAASE, Michaela
The Changing Basis of Economic Responsibility: zur Bedeutung und Rezeption von John Maurice Clarks Artikel zur ökonomischen Verantwortung
Marketing

2017/19 FOSSEN, Frank M.; Ray REES; Davud ROSTAM-AFSCHAR und Viktor STEINER
How Do Entrepreneurial Portfolios Respond to Income Taxation?
Economics

2017/20 NEIDHÖFER, Guido; Joaquín SERRANO und Leonardo GASPARINI
Educational Inequality and Intergenerational Mobility in Latin America: A New Database
Economics

2017/21 SCHMITZ, Sebastian: The Effects of Germany's New Minimum Wage on Employment and Welfare Dependency
Economics

2017/22 WALTER, Paul; Marcus GROß, Timo SCHMID und Nikos TZAVIDIS:
Estimation of Linear and Non-Linear Indicators using Interval Censored Income Data
Economics

2017/23 WAGNER, Julia: Zinsbereinigte Besteuerung und Verlustvortrag : eine Mikrosimulation für deutsche Kapitalgesellschaften
FACTS

2017/24 CRUSIUS, Tobias L. und Marten von WERDER
The Affluency to Quit: How Inheritances Affect Retirement Plannings
Economics

2017/25 ALHO, Juha; Gerrit MÜLLER, Verena PFLIEGER und Ulrich RENDTEL
The Fade Away of an Initial Bias in Longitudinal Surveys
Economics

2017/26 JESSEN, Robin; Maria METZING und Davud ROSTAM-AFSCHAR
Optimal Taxation Under Different Concepts of Justness
Economics

2017/27 LEONE, Tharcisio
The gender gap on intergenerational mobility: Evidence of educational persistence in Brazil
Economics

2017/28 HAAN, Peter; Daniel KEMPTNER und Holger Lüthen
The rising longevity gap by lifetime earnings – distributional implications for the pension system
Economics

2017/29 PREUSS, Malte und Juliane HENNECKE
Biased by Success and Failure: How Unemployment Shapes Stated Locus of Control
Economics