ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)
183, *Part* 1, *pp.* 121–148

Data-driven transformations in small area estimation

Natalia Rojas-Perilla, Sören Pannier and Timo Schmid

Freie Universität Berlin, Germany

and Nikos Tzavidis

University of Southampton, UK

[Received April 2018. Revised May 2019]

Summary. 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 regression model: a feature rarely observed in applications with real data. The paper tackles 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 the linear mixed regression model and transformations are used in small area prediction of linear and non-linear parameters. The use of scaled transformations is crucial as it enables fitting the linear mixed regression 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 by using the parametric and semiparametric (wild) bootstrap. The methods proposed are illustrated by using real survey and census data for estimating income deprivation parameters for municipalities in the Mexican state of Guerrero. Simulation studies and the results from the application show that using carefully selected, data-driven transformations can improve small area estimation.

Keywords: Adaptive transformations; Bootstrap; Maximum likelihood estimation; Poverty mapping; Random effects

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 or 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 income quintile share ratio. Two popular small area methods in this case are the empirical best predictor (EBP), which was proposed by Molina and Rao (2010), and the World Bank method, which was proposed by Elbers *et al.* (2003). Both approaches are based on the use of unit level linear mixed regression models. Although estimation of complex indicators can be implemented also with area level linear mixed regression models (Fabrizi and Trivisano, 2016; Schmid *et al.*, 2017), in this paper we focus on unit level linear mixed regression models. In their original paper, Molina and Rao (2010) assumed that the error terms of the linear mixed regression model follow a Gaussian distribution. In cases where 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? One option is to

Address for correspondence: Timo Schmid, Institut für Statistik und Ökonometrie, Freie Universität Berlin, Garystrasse 21, Berlin 14195, Germany.
E-mail: Timo.Schmid@fu-berlin.de

© 2019 Royal Statistical Society

0964–1998/20/183121

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

formulate the EBP under alternative and more flexible parametric assumptions. Graf *et al.* (2019) studied an EBP method under the generalized beta distribution of the second kind, whereas Diallo and Rao (2014) proposed 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. Another option when the Gaussian assumptions are not satisfied is to use a methodology that minimizes the use of parametric assumptions. For instance, Elbers and van der Weide (2014) proposed an EBP method based on normal mixture models. With this method the distribution of the error terms is described by normal mixtures. Weidenhammer *et al.* (2014) 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.* (2018, 2019). Another option, and the one that 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 in SAE have not been extensively studied or have been studied in a fairly *ad hoc* manner. Elbers *et al.* (2003) and Molina and Rao (2010) suggested the use of logarithmic-type transformations for income data. However, are such transformations the most appropriate choice? Can alternative transformations offer improved estimation? To answer these research questions, the paper investigates data-driven transformations for SAE.

The choice of transformations when modelling income-type outcomes presents different challenges. Transformations should be suitable for dealing with unimodal, leptokurtic and positively skewed data that may include 0 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 transformation as a special case is the Box–Cox family (Box and Cox, 1964). 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 that is described later in Section 4. A solution to this problem can be offered by using the dual power transformation. Although extensive literature on the use of transformations exists (see, for example, John and Draper (1980), Bickel and Doksum (1981) and Yeo and Johnson (2000) among others), in this paper we focus on three types of transformations, namely the 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 SAE 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 (REML) methods for estimating the transformation parameter under the linear mixed regression model following Gurka *et al.* (2006). Alternative estimation approaches based on the minimization of distances (Cramér, 1928; Chakravarti *et al.*,

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

1967) and on the minimization of the skewness (Carroll and Ruppert, 1987) are also discussed.

At this point we should emphasize some of the differences between the present paper and Tzavidis *et al.* (2018). Tzavidis *et al.* (2018) proposes a general framework for the production of small area statistics including measuring uncertainty. Broadly speaking, the framework proposed is based on three stages, namely specification of the problem, analysis of the data and adaptation of the model, and method evaluation. The paper focuses on practical aspects of the SAE process and not on proposing new methodology. The target audience includes practitioners using SAE methods, e.g. colleagues in national statistical institutes. The use of transformations, as a parsimonious approach to adapting the model, is mentioned in the paper but the methodological details are not derived. In contrast, the present paper focuses on developing new and generally applicable methodology that underpins the use of data-driven transformations in SAE and applies the methodology to real data problems. In particular, the current paper proposes the use of the EBP (Molina and Rao, 2010) with data-driven transformations estimated with likelihood-based methods. The paper focuses on scaled transformations that allow the use of standard software for SAE. As we mentioned above, the focus is on the use of the log-shift, Box–Cox, and dual power transformations and the mathematical derivations for developing scaled transformations are presented. Illustrating how to derive scaled transformations for these three transformation types will enable researchers to use similar developments for other families of transformations. In addition, in the present paper we propose two bootstrap schemes (parametric and wild type) for estimating the mean-squared error (MSE) under data-driven transformations and extend these to capture the additional uncertainty due to the estimation of the transformation parameter. The wild bootstrap scheme can been viewed as an insurance policy in case there are some ‘mild’ departures from normality after using transformations. Finally, the present paper includes results from model-based simulation studies that are necessary for comparing the performance of data-driven transformations against the use of fixed, *ad hoc* transformations. Emphasis is given to the estimation of poverty and inequality indicators because of their important socio-economic relevance and policy impact. We further study whether the effect of departures from Gaussian assumptions is different depending on the target of estimation. For instance, departures from normality may have lesser impact on estimates of median income compared with estimate indicators that are more sensitive in the data distribution. The use of model-based simulations was one of the method evaluation approaches that were recommended in Tzavidis *et al.* (2018).

The rest of the paper is structured as follows. The EBP approach is introduced in Section 2. Section 3 presents the survey data that we use in this paper and makes the case, via the use of residual diagnostics, for using transformations. In Section 4 selected transformations are introduced and extended for their use with model-based SAE methods under the 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. In Section 6 the methods proposed are applied to data from Guerrero in Mexico for estimating a range of deprivation and inequality indicators and corresponding estimates of uncertainty. In Section 7 the methods are further evaluated by realistic—for income data—model-based simulations. Section 8 summarizes the main findings and outlines further research.

The programs that were used to analyse the data and a working example can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets.`

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

2. Empirical best prediction method

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 $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^{\mathrm{T}}$ the design matrix containing p explanatory variables and define by s the set of sample units, with s_i the in-sample units in area i and by r 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 $n = \Sigma_{i=1}^{D} n_i$. Hence, we define by $\mathbf{y}_i$ a vector with population elements of the target outcome for area i partitioned as $\mathbf{y}_i^{\mathrm{T}} = (\mathbf{y}_{is}^{\mathrm{T}}, \mathbf{y}_{ir}^{\mathrm{T}})$, where $\mathbf{y}_{is}$ and $\mathbf{y}_{ir}$ denote the sample elements s and the out-of-sample elements r in area i respectively. We now describe in more detail the EBP approach by Molina and Rao (2010), which is the methodology that 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. This was defined by Battese *et al.* (1988) as

$$y_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + u_i + e_{ij}, \qquad u_i \overset{\mathrm{IID}}{\sim} N(0, \sigma_u^2), \quad e_{ij} \overset{\mathrm{IID}}{\sim} N(0, \sigma_e^2), \tag{1}$$

where u_i, the area-specific random effects, and e_{ij}, the unit level error, are assumed to be independent. Assuming normality for the unit level error and the area-specific random effects, the conditional distribution of the out-of-sample data given the sample data is also normal. A Monte Carlo approach is used to obtain a numerically efficient approximation to the expected value of this conditional distribution as follows.

Step 1: use the sample data to obtain $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$ and the weighting factors $\hat{\gamma}_i = \hat{\sigma}_u^2 / (\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_i)$.
Step 2: for $l = 1, \ldots, L$,

(a) generate $v_i^{(l)} \sim^{\mathrm{IID}} N\{0, \hat{\sigma}_u^2(1 - \hat{\gamma}_i)\}$ and $e_{ij}^{(l)} \sim^{\mathrm{IID}} N\{0, \hat{\sigma}_e^2\}$ and obtain a pseudopopulation of the target variable by

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

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.

Step 3: finally, take the mean over the L Monte Carlo runs in each area to obtain a point estimate of the indicator of interest,

$$\hat{I}_i^{\mathrm{EBP}} = \frac{1}{L} \sum_{l=1}^{L} I_i^{(l)}.$$

As is common in real applications, some areas are out of sample. For those areas, we cannot estimate an area-specific random effect, and hence the corresponding area-specific random effect is set equal to 0. Synthetic values of the outcome for the out-of-sample areas are then generated under the linear mixed regression model as follows:

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

with $u_i^{(l)} \sim^{\mathrm{IID}} N(0, \hat{\sigma}_u^2)$ and $e_{ij}^{(l)} \sim^{\mathrm{IID}} N(0, \hat{\sigma}_e^2)$. Finally, a parametric bootstrap—under the model assumed—is used for the MSE estimation. This is discussed in detail in Section 5. Assuming normality for the error terms is a convenient assumption as it enables the conditional distri-

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

bution of $\mathbf{y}_r|\mathbf{y}_s$ to be derived. However, in applications that involve modelling an income-type outcome, as in this paper, assuming normality is unrealistic. If our primary target is to develop a methodology that can easily be used in practice, finding appropriate data transformations is important.

3. The Guerrero case-study: source of data and initial analysis

In this section, we describe the sources of data that were used in the application and provide a motivation for the use of transformations. The case-study was carried out by using the open-source software R (R Core Team, 2017) and R packages. The data that we use in this paper come from the Mexican state of Guerrero: one of the 32 states in Mexico. The state Guerrero is considered by the World Bank to be one of the states—next to the State of Mexico that was investigated by Tzavidis *et al.* (2018)—mostly contributing to income inequality in Mexico (Bedoya *et al.*, 2013). Additionally, according to the United Nations Development Programme (UNDP), Guerrero has one of the highest rates of poverty and lack of infrastructural development (Tortajada, 2006). According to the general social development law in Mexico, the National Institute of Statistics and Geography (the Instituto Nacional de Estadística y Geografía) must provide relevant official statistics at the national, state and municipal levels. Furthermore, the Social Development Law (the *Ley General de Desarrollo Social*) in Mexico establishes that the National Council for the Evaluation of Social Development Policy (the Consejo Nacional de Evaluación de la Política de Desarrollo Social) should measure poverty at state level every 2 years and at municipal level every 5 years. For carrying out the analysis the statistical and geographical information was provided by the Instituto Nacional de Estadística y Geografía through the Household Income and Expenditure Survey (the *Encuesta Nacional de Ingresos y Gastos de los Hogares* (ENIGH)) 2010 and the National Population and Housing Census of 2010. Looking in more detail at the data that were available and their geographic coverage, Guerrero comprises 81 administrative divisions, known as municipalities. From the 81 municipalities 40 municipalities with 1611 households are in sample (in the sample of the ENIGH survey) and the remaining 41 municipalities are out of sample. For the in-sample municipalities the maximum sample size in a municipality is 511, the minimum is 9 and the median is 24 households. Note that more than 30% of the sample is from a single municipality, the capital (Chilpancingo de los Bravo).

The survey and census data include a large number of sociodemographic variables, which are common and are measured similarly in both sources of data. The total household *per capita* income (*ictpc*, measured in pesos) is a variable that is recorded for households and is available in the survey but not in the census. We used this variable as a proxy that best approximates the living standard in Guerrero and as the outcome variable in our models. Socio-economic variables that are available for the households in both the survey and the census data are used as explanatory variables. The underlying linear mixed regression model (1) of the EBP has two levels: households and municipalities. The variables that are available in the survey and census data, which are identified by using the Bayesian information criterion (BIC) as good predictors of ictpc, are described in Table 1. From now on, the working model is assumed to be known and fixed.

The next step after the identification of a possible set of covariates is assessing the predictive power of the model. Nakagawa and Schielzeth (2013) proposed the use of two coefficients of determination that are suitable for linear mixed 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 the random effects.

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

Table 1. Description of the explanatory variables used in the working model

Determinant	*Variable*
Occupation	1: indicator of whether 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
Socio-economic 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

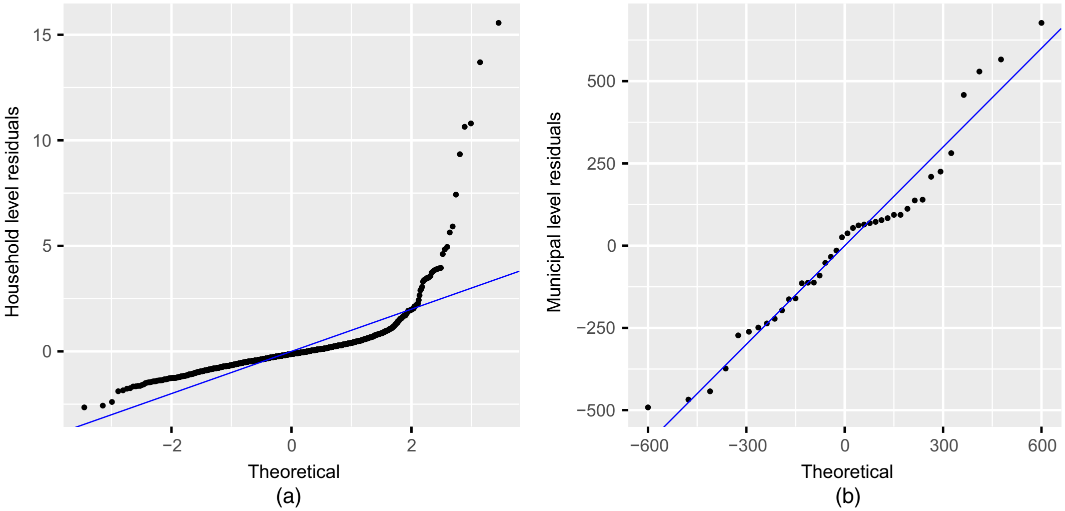


Fig. 1. *Q*–*Q*-plots of the (a) household and (b) municipal level error terms

Without using any transformation, these measures are both around 35% and the corresponding intraclass correlation ICC under the model is 0.027.

To explore the validity of the Gaussian assumptions underlying the linear mixed regression model, it is common practice to perform normality tests and some residual diagnostics. The p-values of the Shapiro–Wilk test statistic are equal to 2.2×10^{-16} for the household level and 0.197 for the municipal level. These results indicate that the null hypothesis of normality for the household level is rejected. As normality tests like the Shapiro–Wilk test have some problems we also present some visual approaches in addition. Fig. 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 ictpc-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 0 and 3 respectively. The skewness and kurtosis of the household level are equal to 6.338 and 75.483, and for the municipal level equal to 0.448 and 3.250. These results indicate severe departures—especially for

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

the household level—from the Gaussian assumptions when modelling the non-transformed income.

4. Use of transformations

To approach closer to normality, it is common to use a one-to-one transformation $T(y_{ij}) = y^*_{ij}$ of the target variable. The application of the natural logarithmic transformation, which is a popular choice for income data, leads in many cases from right-skewed to more symmetric distributions. This approach was followed by Molina and Martín (2018), in which the logarithmic transformation was applied to an income-type variable for meeting the assumptions of the model that was proposed by Battese *et al.* (1988). In particular, Molina and Martín (2018) proposed analytic MSE estimators and developed bias correction terms that are necessary when estimating small area averages by using a logarithmic transformation under the linear mixed regression model. The logarithmic transformation is frequently used for dealing with non-normality because of its simplicity. However, can an alternative transformation with data-driven parameter(s) λ, $T_\lambda(y_{ij}) = y^*_{ij}(\lambda)$, offer small area estimates with improved precision?

The structure of the section is as follows. In Section 4.1 we introduce the EBP approach with data-driven transformations. In Section 4.2 we propose likelihood-based approaches for estimating the transformation parameter λ in general and discuss three particular subcases—the log-shift, Box–Cox, and dual power transformations—in detail. Finally, in Section 4.3 we discuss alternatives to likelihood-based approaches for estimating the transformation parameter.

4.1. Empirical best predictor under transformations

To apply the EBP method by using transformations, the linear mixed regression model is redefined as follows:

$$y^*_{ij}(\lambda) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + u_i + e_{ij}, \qquad u_i \overset{\mathrm{IID}}{\sim} N(0, \sigma_u^2) \text{ and } e_{ij} \overset{\mathrm{IID}}{\sim} N(0, \sigma_e^2). \tag{2}$$

The EBP approach under transformations can be rewritten as follows.

Step 1: select a transformation and obtain $T_\lambda(y_{ij}) = y^*_{ij}(\lambda)$.

Step 2: use the transformed sample data to obtain $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ and calculate the weighting factors $\hat{\gamma}_i = \hat{\sigma}_u^2 / (\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_i)$.

Step 3: for $l = 1, \ldots, L$,

(a) generate $v_i^{(l)} \sim^{\mathrm{IID}} N\{0, \hat{\sigma}_u^2(1 - \hat{\gamma}_i)\}$ and $e_{ij}^{(l)} \sim^{\mathrm{IID}} N(0, \hat{\sigma}_e^2)$ and obtain a pseudopopulation of the target variable by

$$y_{ij}^{*(l)} = \mathbf{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \hat{u}_i + v_i^{(l)} + e_{ij}^{(l)};$$

(b) back-transform $y_{ij}^{*(l)}$ to the original scale $y_{ij}^{(l)} = T_\lambda^{-1}(y_{ij}^{*(l)})$;

(c) calculate the indicator of interest $I_i^{(l)}$ in each area.

Step 4: finally, take the mean over the L Monte Carlo generations in each area to obtain an estimate of the indicator of interest,

$$\hat{I}_i^{\mathrm{EBP}} = \frac{1}{L} \sum_{l=1}^{L} I_i^{(l)}.$$

4.2. Likelihood-based approach for estimating λ

For estimating the transformation parameter λ, the linear mixed regression model defined in

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

expression (2) is used. Assume that the transformed vectors $\mathbf{y}_i^*$ are independent and normally distributed for some unknown λ,

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

where

$$\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$$

and

$$\mathbf{V}_i = \sigma_u^2 \mathbf{1}_{N_i}\mathbf{1}'_{N_i} + \sigma_e^2 \mathbf{I}_{N_i},$$

with $\mathbf{1}_{N_i}$ a column vector of 1s of size N_i and $\mathbf{I}_{N_i}$ the $N_i \times N_i$ identity matrix; the vector of unknown model parameters is $\boldsymbol{\theta}^{\mathrm{T}} = (\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2, \lambda)$. The log-likelihood function under the model is defined as follows:

$$l_{\mathrm{ML}}(\mathbf{y}^*, \lambda|\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{D}\log|\mathbf{V}_i| - \frac{1}{2}\sum_{i=1}^{D}(\mathbf{y}_i^*(\lambda) - \mathbf{X}_i\hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{y}_i^*(\lambda) - \mathbf{X}_i\hat{\boldsymbol{\beta}}).$$

The log-likelihood function in relation to the original observations is obtained by multiplying the normal density by the logarithm of the Jacobian of the transformation from $\mathbf{y}_i$ to $\mathbf{y}_i^*(\lambda)$. The Jacobian $J(\lambda, \mathbf{y})$ is defined as

$$\prod_{i=1}^{D}\prod_{j=1}^{n_i}\left|\frac{\mathrm{d}y_{ij}^*(\lambda)}{\mathrm{d}y_{ij}}\right|$$

and is incorporated as follows:

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

The maximization of $l_{\mathrm{ML}}(\mathbf{y}, \lambda|\boldsymbol{\theta})$ produces maximum likelihood 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, REML theory is recommended (Verbeke and Molenberghs, 2000). REML is defined as follows:

$$\begin{aligned} l_{\mathrm{REML}}(\mathbf{y}, \lambda|\boldsymbol{\theta}) = &-\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\mathbf{V}_i| - \frac{1}{2}\log\left|\sum_{i=1}^{D}\mathbf{X}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}(\mathbf{y}_i^*(\lambda) - \mathbf{X}_i\hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{y}_i^*(\lambda) - \mathbf{X}_i\hat{\boldsymbol{\beta}}) + \log\{J(\lambda, \mathbf{y})\}. \end{aligned} \tag{3}$$

The use of the scaled version of a selected transformation, defined by $y_{ij}^*(\lambda)/J(\lambda, \mathbf{y})^{1/n} = z_{ij}^*(\lambda)$, is crucial for estimating the transformation parameter λ under the REML approach that was presented above. The Jacobian of such a scaled transformation is equal to 1. This means that the scale of the likelihood is preserved independently of the transformation and its parameter λ. Therefore, values of the log-likelihood function—under differently transformed $y_{ij}^*(\lambda)$—can be directly compared and the log-likelihood function simplifies to the log-likelihood function of the linear mixed regression model. As a result, standard software for fitting this model can be used to estimate the transformation parameter λ. Even though using scaled transformations aids

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

Table 2. Jacobian and scaled data-driven transformations for log-shift, Box–Cox and dual transformations

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

the implementation of the methods in practice, appropriate scaling factors must be developed depending on the type of transformation that is used.

Although the theory is applicable to data-driven transformations in general, we focus on the three types of transformations that we presented in Section 1, namely the log-shift, Box–Cox and dual power transformations. Additionally, we use the frequently applied logarithmic transformation as a benchmark. This transformation is defined by $y^*_{ij}=\log(y_{ij}+s)$, where s denotes a fixed parameter such that $y_{ij}+s>0$. The log-shift transformation (Yang, 1995), presented below, extends the logarithmic transformation by including the data-driven transformation parameter $\lambda\geqslant s$ which needs to be estimated:

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

When $\lambda=s$, the logarithmic transformation is obtained. The Box–Cox transformation (Box and Cox, 1964) is defined as follows:

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

where s denotes a fixed parameter such that $y_{ij}+s>0$. If $\lambda=0$, the logarithmic transformation is then a special case and, if $\lambda=1$, the data are only shifted. One difficulty with Box–Cox-type transformations is the long-standing truncation, i.e. $y^*_{ij}(\lambda)$ is bounded, from below by $1/\lambda$ if $\lambda>0$ and from above by $-1/\lambda$ if $\lambda<0$. This is the key motivation for the third type of transformation. The dual power transformation, which was introduced by Yang (2006), is defined as follows:

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

where s is defined as in the case of the Box–Cox transformation.

The corresponding Jacobian that is used in equation (3) and scaled versions of the log-shift, Box–Cox and dual power transformations are presented in Table 2. For more details we refer to the developments in Appendix A.

4.3. Alternative approaches for estimating λ

The maximum likelihood and REML approaches that were introduced in Section 4.2 rely on

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

parametric assumptions that may be influenced by outliers in the data. The kurtosis and skewness are crucial features for defining the shape of a distribution and a proximity measure can be minimized to find a transformation parameter under which the empirical distribution of residuals has skewness and kurtosis as close as possible to 0 and 3 respectively. In general, skewness is considered more important than kurtosis; therefore, minimizing the skewness is an approach that has already been considered in the literature (Royston and Lambert, 2011) for linear models as follows:

$$\hat{\lambda}_{\text{skew}} = \arg\min_{\lambda} |S_{e_\lambda}|,$$

where S_{e_λ} is the skewness and $\sigma^2_{e_\lambda}$ denotes the variance of the unit level error terms. Note that the index λ is used to emphasize that the skewness and the variance parameters depend on the transformation parameter. 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 that uses a weight w to ensure that, the larger the error term variance $\sigma^2_{e_\lambda}$ is, the more weight its skewness will have in the minimization. Let S_{u_λ} be the skewness and $\sigma^2_{u_\lambda}$ be the variance of the area-specific random effects u_i of the linear mixed regression model. The estimation criterion in the pooled skewness approach is defined as follows:

$$\hat{\lambda}_{\text{poolskew}} = \arg\min_{\lambda} \{w|S_{e_\lambda}| + (1-w)|S_{u_\lambda}|\},$$

where

$$w = \frac{\hat{\sigma}^2_{e_\lambda}}{\hat{\sigma}^2_{u_\lambda} + \hat{\sigma}^2_{e_\lambda}}.$$

Considering only the skewness may ignore other properties of the distribution. Hence, a measure describing the distance between two distribution functions is another alternative. Two distance measures, the Kolmogorov–Smirnov and the Cramér–von Mises statistics KS and CvM respectively are used:

$$\hat{\lambda}_{\text{KS}} = \arg\min_{\lambda} \sup|F_n(\cdot) - \Phi(\cdot)|,$$

$$\hat{\lambda}_{\text{CvM}} = \arg\min_{\lambda} \int_{-\infty}^{\infty} \{F_n(\cdot) - \Phi(\cdot)\}^2 \phi(\cdot),$$

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. The effect of using alternative approaches for estimating λ is studied in a model-based simulation study in Section 7.3.

5. Mean-squared error estimation under transformations

Estimating the MSE of small area estimates is a challenging problem. In the case of the EBP Molina and Rao (2010) proposed a parametric bootstrap procedure following González-Manteiga *et al.* (2008). In this section we propose two bootstrap schemes for estimating the MSE under transformations. These bootstrap MSE estimators are extended to capture the additional uncertainty due to the estimation of the transformation parameter λ. The difference between the two bootstrap schemes is the mechanism that is used for generating the bootstrap population. In particular, the first bootstrap generates bootstrap realizations of the random effects

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

and unit level error terms parametrically. In contrast, the second bootstrap is a semiparametric wild bootstrap which aims to protect against departures from the assumptions of the model, in particular, those of the unit level error term.

The steps of the proposed parametric bootstrap are as follows.

Step 1: for $b=1,\ldots,B$,

(a) using the sample estimates, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$, $\hat{\sigma}_e^2$ and $\hat{\lambda}$, generate $u_i^{(b)} \sim^{\text{IID}} N(0,\hat{\sigma}_u^2)$ and $e_{ij}^{(b)} \sim^{\text{IID}} N(0,\hat{\sigma}_e^2)$ and simulate a bootstrap superpopulation $y_{ij}^{*(b)} = \mathbf{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}} + u_i^{(b)} + e_{ij}^{(b)}$;

(b) back-transform $y_{ij}^{*(b)}$ to the original scale $y_{ij}^{(b)} = T_{\hat{\lambda}}^{-1}(y_{ij}^{*(b)})$ 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 (as the back-transformed sample data are used, the transformation parameter λ is re-estimated in each bootstrap replication b);

(d) obtain $\hat{I}_{i,b}^{\text{EBP}}$.

Step 2: $\widehat{\text{MSE}}(\hat{I}_i^{\text{EBP}}) = B^{-1}\Sigma_{b=1}^{B}(\hat{I}_{i,b}^{\text{EBP}} - I_{i,b})^2$.

As mentioned before, the proposed parametric bootstrap allows for the additional uncertainty due to the estimation of the transformation parameter. Although the use of an optimal transformation may reduce the deviation from normality, there may still be departures from normality especially in the tails of the distribution of the unit level error term. To overcome this problem, we propose a semiparametric bootstrap that relies on the normality of the random effects but generates the unit level error terms by using the empirical distribution of suitably scaled unit level residuals. The proposed wild bootstrap scheme is described below.

Step 1: fit model 1 by using an appropriate transformation $T(y_{ij}) = y_{ij}^*$ and obtain $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$, $\hat{\sigma}_e^2$ and $\hat{\lambda}$.

Step 2: calculate the sample residuals by $\hat{e}_{ij} = y_{ij} - \mathbf{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \hat{u}_i$.

Step 3: scale and centre the residuals by using $\hat{\sigma}_e$. The scaled and centred residuals are denoted by $\hat{\epsilon}_{ij}$.

Step 4: for $b=1,\ldots,B$,

(a) generate $u_i^{(b)} \sim^{\text{IID}} N(0,\hat{\sigma}_u^2)$;

(b) calculate the linear predictor $\eta_{ij}^{(b)}$ by $\eta_{ij}^{(b)} = \mathbf{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}} + u_i^{(b)}$;

(c) match $\eta_{ij}^{(b)}$ with the set of estimated linear predictors $\{\hat{\eta}_k | \eta \in n\}$ from the sample by using

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

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

(d) generate weights w from a distribution satisfying the conditions in Feng *et al.* (2011) where w is a simple two-point mass distribution with probabilities 0.5 at $w=1$ and $w=-1$;

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

(f) back-transform $T(y_{ij}^{*(b)})$ to the original scale and compute the population value $I_{i,b}$;

(g) extract the bootstrap sample in $y_{ij}^{(b)}$ and use the EBP method, as described in Section 4;

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

Step 5: $\widehat{\text{MSE}}_{\text{wild}}(\hat{I}_i^{\text{EBP}}) = B^{-1}\Sigma_{b=1}^{B}(\hat{I}_{i,b}^{\text{EBP}} - I_{i,b})^2$.

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

The performance of both MSE estimators is compared in a model-based simulation study in Section 7.

6. The Guerrero case-study: application of data-driven transformations

The benefits of using the proposed EBP approach with data-driven transformations for estimating deprivation and inequality indicators are illustrated in an application using the household data from the ENIGH survey 2010 and the National Population and Housing Census 2010 that we introduced in Section 3. The aim is to estimate the head count ratio HCR and the poverty gap PGAP as well as the income quintile share ratio QSR for the 81 municipalities in Guerrero. As the ENIGH survey and the census contain information only on household level we estimate the poverty and inequality indicators for households and not individuals.

The indicators HCR and PGAP are special cases of the Foster–Greer–Thorbecke indicators (Foster *et al.*, 1984) and they depend on a poverty line t which is equal to 0.6 times the median of the target variable. The Foster–Greer–Thorbeke index of type α for an area i is defined by

$$F_i(\alpha,t)=\frac{1}{N_i}\sum_{j=1}^{N_i}\left(\frac{t-y_{ij}}{t}\right)^{\alpha}\mathbb{I}(y_{ij}\leqslant t),\qquad \text{for } \alpha=0,1,2,$$

where $\mathbb{I}(\cdot)$ denotes an indicator function which returns 1 if '$(\cdot)$' holds and 0 otherwise. When $\alpha=0$, $F_i(\alpha,t)$ is the head count ratio and represents the proportion of the households whose income is below the poverty line t. Taking $\alpha=1$, $F_i(\alpha,t)$ defines PGAP, which is a measure of poverty intensity and quantifies the degree to which the average income of people living under the poverty line differs from the poverty line. In addition to the two deprivation indicators, we investigate inequality by the quintile share ratio defined by

$$\mathrm{QSR}_i=\frac{\sum_{j=1}^{N_i}\mathbb{I}(y_{ij}\geqslant \mathbf{y}_{0.8})y_{ij}}{\sum_{j=1}^{N_i}\mathbb{I}(y_{ij}\leqslant \mathbf{y}_{0.2})y_{ij}},$$

where $\mathbf{y}_{0.8}$ and $\mathbf{y}_{0.2}$ denote the 80% and 20% quantiles of the target variable respectively. QSR is a widely used inequality indicator because of its simplicity and straightforward interpretation (Eurostat, 2004). The estimation of QSR is challenging, characterized by a large variability also in large samples. However, we have decided to report the estimated QSR in this paper only for illustration. In particular, we are interested in showing the increasing importance of the model assumptions when the target parameter depends of the tails of the data distribution and the importance of using data transformation parameters.

Before focusing on the state of Guerrero, we briefly illustrate the need for data-driven transformations in different states in Mexico. Fig. 2 represents the estimated data-driven Box–Cox transformation parameters $\hat{\lambda}$ (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. Furthermore, we observe that a fixed logarithmic transformation is not suitable for any of the states.

6.1. Model checking and residual diagnostics

In Section 3 we show that the model assumptions of the working model in the state of Guerrero are not met. We now discuss the use of the proposed data-driven transformations for adapting

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

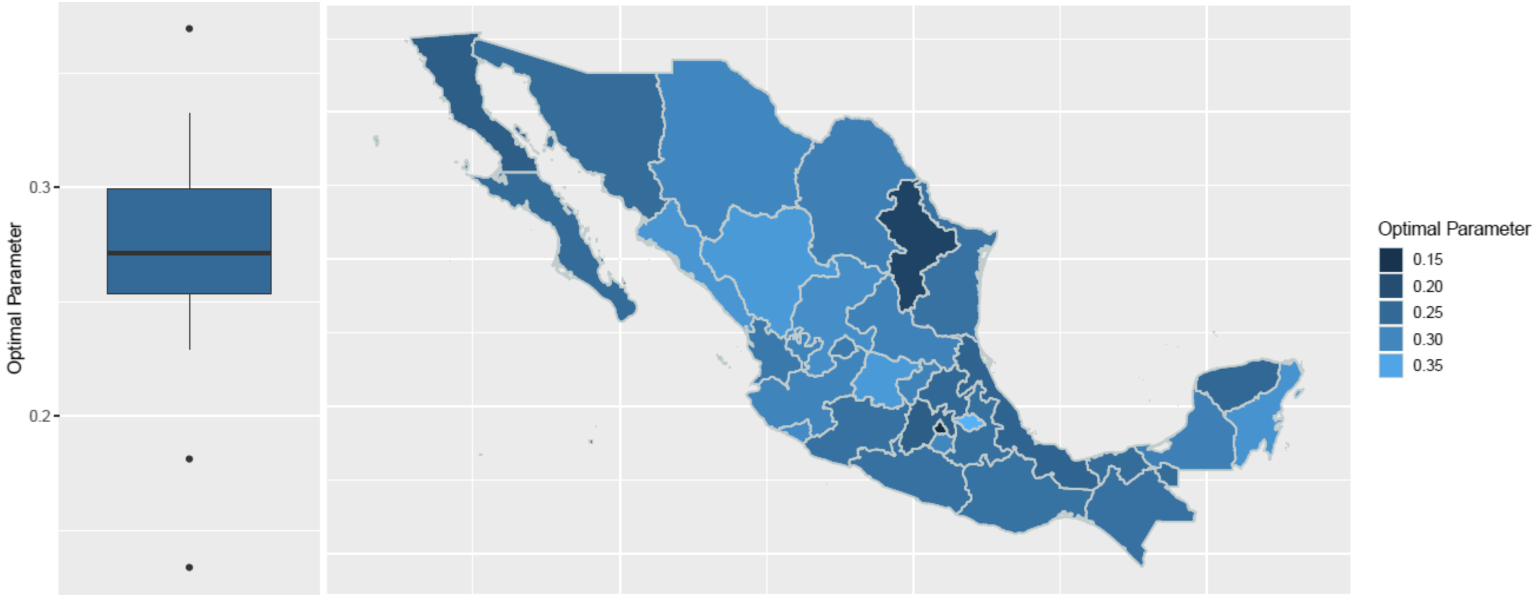
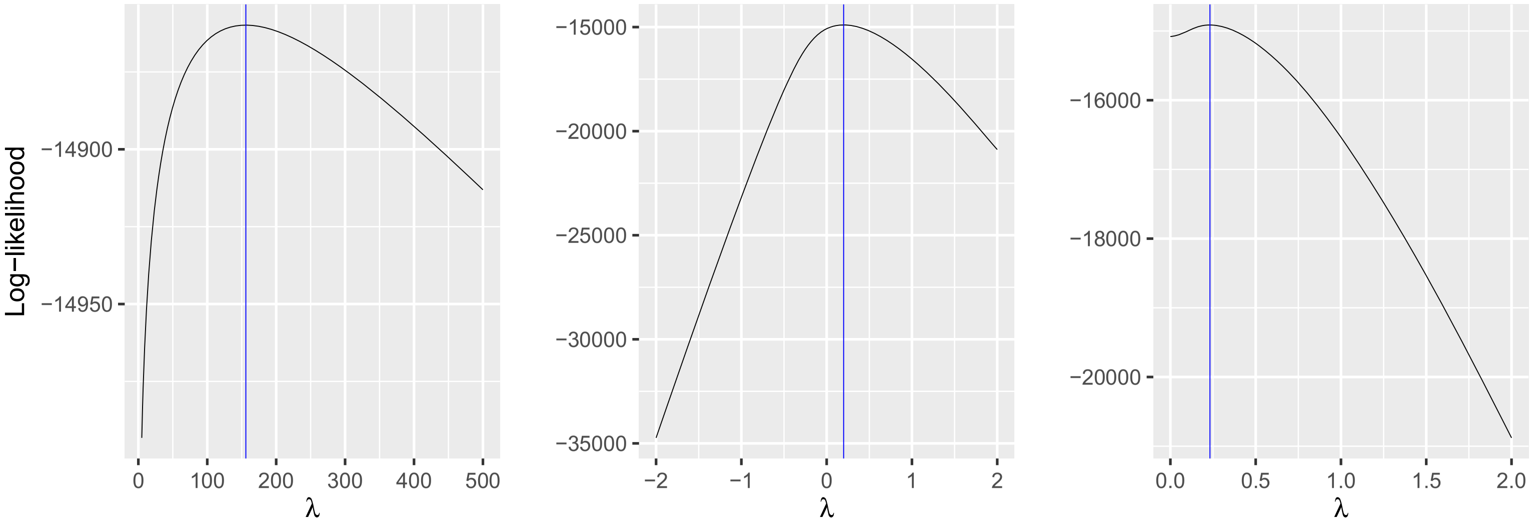


Fig. 2. Estimated transformation parameters of the Box–Cox transformation in the various states of Mexico

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025



(a) (b) (c)

Fig. 3. Optimal transformation parameter λs for the working model under the (a) log-shift, (b) Box–Cox and (c) dual power transformations in the state of Guerrero

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

Table 3. R^2_{m}, R^2_{c}, λs and ICC for the working model under the various transformations

Transformation	R^2_m	R^2_c	λ	*ICC*
No	0.351	0.368	—	0.027
Log	0.361	0.458	—	0.151
Log-shift	0.460	0.522	156.443	0.114
Box–Cox	0.454	0.513	0.199	0.108
Dual	0.454	0.512	0.232	0.106

Table 4. Skewness, kurtosis and values of the Shapiro–Wilk *p*-values for the municipal and household level error terms of the working models for the EBP under the various transformations

Transformation	*Household level residuals*			*Municipal level residuals*		
	Skewness	*Kurtosis*	*p-value*	*Skewness*	*Kurtosis*	*p-value*
No	6.338	75.483	0.000	0.448	3.250	0.197
Log	−2.046	16.986	0.000	−1.491	7.059	0.001
Log-shift	−0.024	4.143	0.000	−0.276	3.485	0.893
Box–Cox	−0.055	6.085	0.000	−0.389	3.861	0.662
Dual	−0.045	6.542	0.000	−0.387	3.889	0.657

the working model. In particular, we focus on the three data-driven transformations that were presented in Section 4.2, denoted by *log-shift*, *Box–Cox* and *dual* power transformations and their comparison with

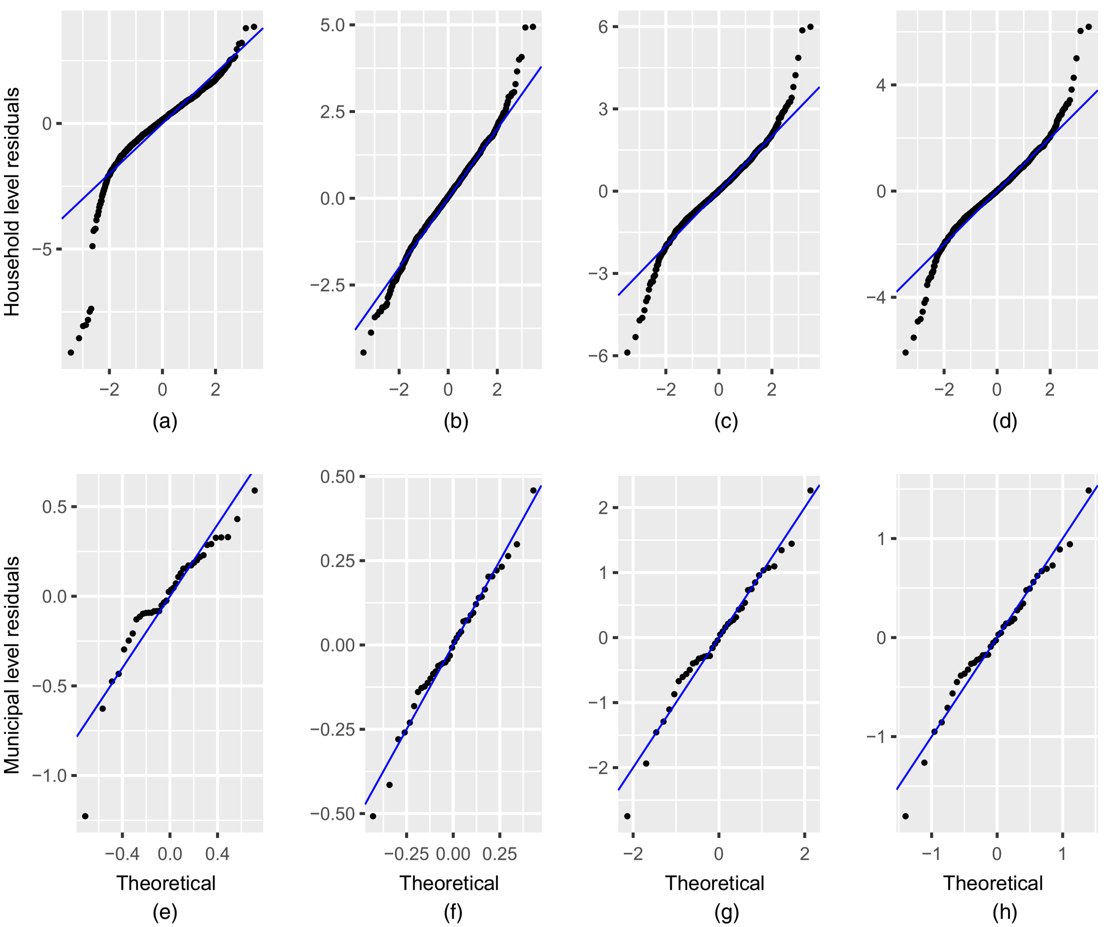
(a) a model that uses a logarithmic transformation (*Log*) and
(b) a model that uses the untransformed income variable (*No*).

To start with, Fig. 3 provides a graphical representation of the REML maximization for the transformation parameter λ for the log-shift, Box–Cox and dual power transformations in the state of Guerrero. In this case the optimal λs are approximately equal to 156.44, 0.20 and 0.23 respectively (Table 3).

To analyse whether the use of transformations improves the predictive power of the model, Table 3 reports the percentage of variability explained for each model and its corresponding ICC. As ICC is larger than 0 in all cases, there appears to be unexplained between-area variability and hence the use of the mixed model may be appropriate. Using the untransformed ictpc-outcome leads to marginal (R^2_{m}) and conditional (R^2_{c}) coefficients of determination of 0.35 and 0.37 respectively. The use of a logarithmic transformation improves the predictive power of the model in terms of the conditional R^2_{c} and the marginal R^2_{m}. However, it can clearly be noted that the use of data-driven transformations increases the predictive power of the model.

A detailed analysis of the Gaussian assumptions of the working models corresponding to each transformation is now carried out. The results summarizing the skewness, kurtosis and Shapiro–Wilk normality tests are presented in Table 4 and the *Q*–*Q*- plots are presented in Fig. 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

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 4. *Q*–*Q*-plots for (a)–(d) the Pearson household level and (e)–(h) municipal level residuals of the working model for the EBP under the various transformations: (a), (e) log-transformation; (b), (f) log-shift transformation; (c), (g) Box–Cox transformation; (d), (h) dual transformation

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

Table 5. Summaries of point estimates and corresponding RMSEs over municipalities in Guerrero

Transformation	*HCR*		*PGAP*		*QSR*	
	Mean	*Median*	*Mean*	*Median*	*Mean*	*Median*
Point estimation						
Log	0.48	0.49	0.24	0.23	18.03	17.82
Log-shift	0.44	0.44	0.21	0.20	15.56	14.39
Box–Cox	0.44	0.44	0.22	0.22	16.78	16.39
Dual	0.44	0.44	0.22	0.22	17.38	16.98
RMSE						
Log	0.11	0.13	0.07	0.08	32.19	27.54
Log-shift	0.09	0.09	0.06	0.05	4.68	1.91
Box–Cox	0.09	0.10	0.06	0.06	2.98	2.95
Dual	0.09	0.10	0.06	0.06	2.85	2.76

transformation and without a transformation. The picture considerably improves for the data-driven transformations. The log-shift, Box–Cox and dual power transformations lead to very similar results in terms of skewness and kurtosis. We note that the log-shift transformation performs slightly better in terms of kurtosis and skewness compared with the Box–Cox and dual power transformation. These findings are supported by the Q–Q-plots that are displayed in Fig. 4.

The data-driven transformations lead to similar Q–Q-plots with more symmetrical and less extreme tails compared with the fixed log-transformation. 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 with the use of a fixed logarithmic transformation.

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 that were 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 that we introduced in Section 5 with $B=500$ bootstrap replications.

Table 5 shows summaries over municipalities of point estimates and root MSEs (RMSEs) under the various transformations. In addition we provide a detailed comparison between EBP methods under the transformations and the direct estimator with corresponding coefficients of variation as part of the on-line supplementary material. We observe that the estimates based on the EBP with data-driven transformations are more efficient on average (in terms of RMSE) than the corresponding estimates based on a fixed logarithmic transformation. The effect is especially pronounced for indicators that rely on the tail of the distribution like QSR. Furthermore, the use of data-driven transformations also has an effect on the point estimates of the indicators. For HCR and PGAP, the results that were obtained under the three data-driven transformations are similar to each other; main differences are noticeable when just the fixed logarithmic transformation and no transformation are applied. For instance, the EBP estimates under the model with a logarithmic transformation are on average 5% higher compared with the EBP estimates with data-driven transformations for HCR (see Table 1 in the on-line supplementary material).

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

In addition, the distribution (over municipalities) of the point estimates that are obtained under the EBP with data-driven transformations appear to be closer to the distribution of the direct estimates than the distribution of the EBP under a fixed logarithmic transformation.

Having assessed the estimates from a statistical perspective, we investigate the results in the context of the spatial distribution of poverty and inequality in the state of Guerrero. Fig. 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 show the results only for the EBP with the log-shift transformation. We observe clear regional differences between the municipalities. Having a closer look at the coastal area in the south-west of Guerrero, where the largest city Acapulco is located, we observe lower levels of poverty (HCR and PGAP) and inequality (QSR) compared with other parts of the state. The coastline to the Pacific Ocean is wealthier because of several tourist destinations like Acapulco, Ixtapa and Zihuatanejo. In contrast, there is also a clear deprivation hotspot in the eastern part of the state of Guerrero (e.g. municipalities: Cochoapa el Grande, Metlatnoc and Atlamajalcingo del Monte) with high poverty and inequality rates. These municipalities are home to indigenous populations living in isolated mountain areas.

7. Model-based simulation study

In this section, we present results from a model-based simulation study that aims to evaluate the performance of the methods proposed. In Section 7.1 we analyse the behaviour of the data-driven transformation parameter under four scenarios for the distributions of the area and unit level error terms. In Section 7.2 we investigate the ability of the proposed methods to provide more precise small area estimates than the EBP with a fixed logarithmic transformation or without a transformation and assess the performance of the proposed MSE estimators. Finally, in Section 7.3 we evaluate the methods for estimating the transformation parameter. In addition we also conducted a design-based simulation study with a variable available in the census data that is highly correlated with the target variable (ictpc) in the application. The results are provided as part of the on-line supplementary material.

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 stratified random sampling with strata defined by the 50 small areas. This leads to a sample size of $n=\Sigma_{i=1}^{D} n_i=921$ whereby the area-specific sample sizes n_i vary between 8 and 29. We chose the sample sizes mainly for two reasons. First, we want to assess the data-driven transformations under extreme but realistic cases. Second, the sample sizes are similar in the case-study.

Four scenarios, which are denoted by *Normal*, *Log-scale*, *Pareto* and *GB2* (generalized beta distribution of the second kind), are considered. Details about the data-generating mechanisms of the scenarios are provided in Table 6. Under scenario Normal, data are generated by using

Table 6. Model-based simulation settings for the analysis of the MSE

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

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

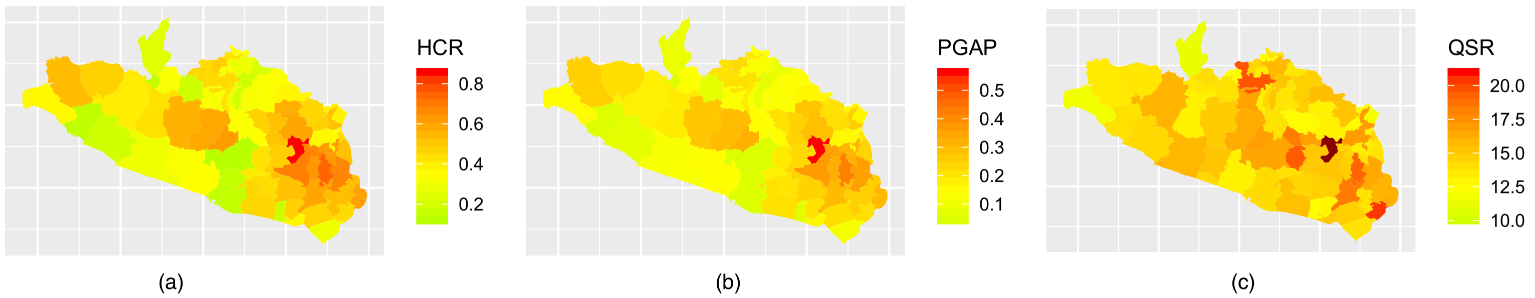


Fig. 5. Maps of (a) HCR, (b) PGAP and (c) QSR in Guerrero for the EBP method under the log-shift transformation at municipal level

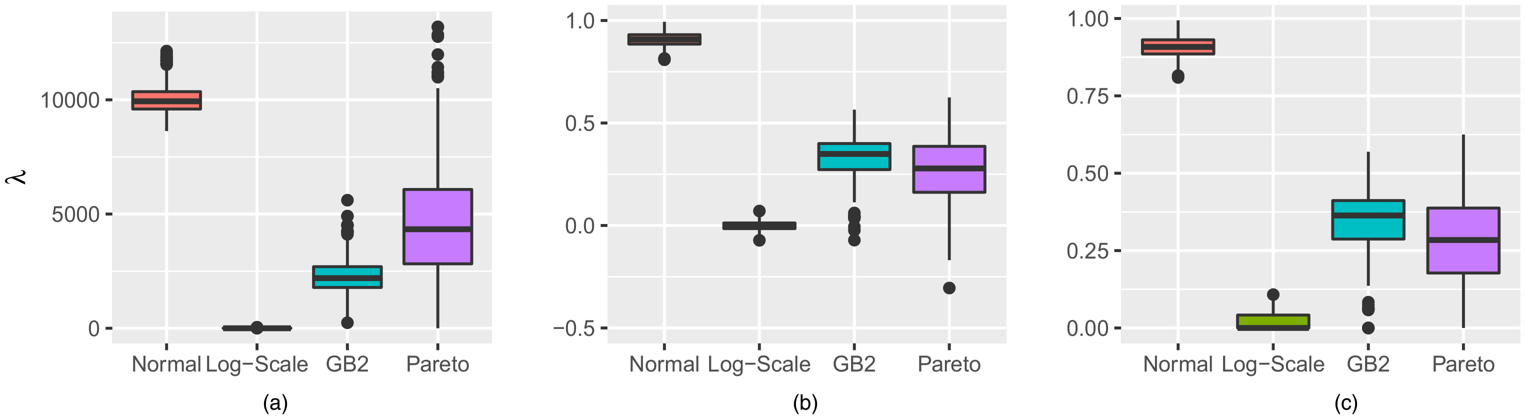


Fig. 6. Estimated transformation parameters for (a) the log-shift, (b) Box–Cox and (c) dual power transformations under the various settings

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

normal distributions for the random effects and unit level errors. Under the second scenario, random effects and unit level errors are generated under a log-normal distribution such that a fixed logarithmic transformation is suitable. Scenarios Pareto and GB2 are settings that attempt to replicate realistic situations for income data. In particular, these distributions mimic the features of income-based variables, namely unimodal, leptokurtic and highly skewed data influenced by outliers. Random effects are generated by using a normal distribution and unit level error terms are generated under a Pareto and GB2 scenario respectively. Each setting was repeated independently $M=500$ times. We focus on the three data-driven transformations, namely log-shift, Box–Cox and dual power transformations, and compare these with the case of a fixed logarithmic transformation and the case of using untransformed data.

7.1. Behaviour of the data-driven transformation parameters

Fig. 6 shows boxplots of the estimated transformation parameters λ for the log-shift, Box–Cox and dual power transformations (over $M=500$ replications) under the four simulation settings. The data-driven transformation parameters are estimated by REML. Under the Normal setting the parameters of the Box–Cox and dual power transformations are close to 1, indicating that no transformation is needed. In the Log-scale scenario, the data were generated in such a way that normality may be achieved by applying the logarithmic transformation. In this case the log-shift transformation parameter is close to 0 and the same holds for the parameters of the Box–Cox and dual power transformations. For the other two scenarios (Pareto and GB2), the data-driven parameters are between 0.25 and 0.5, so neither using a logarithmic transformation nor ignoring the need for a transformation is appropriate. Overall, the results indicate that the data-driven transformations behave as expected in the four scenarios and adapt to the shapes of the data distributions.

7.2. Performance of the empirical best predictor under data-driven transformations

In this section we compare the performance of the proposed methods with the case of

(a) fixed logarithmic transformation and
(b) no transformation.

We then assess the performance of the MSE estimators. Five estimators of small area deprivation and inequality indicators (HCR, PGAP and QSR) are evaluated. The EBP and the corresponding MSE estimators are implemented by using $L=100$ and $B=500$. The following quality measures averaged over Monte Carlo replications M are used to assess the performance of a small area estimator in area i:

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

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

where $\hat{I}_i^{\text{method}}$ denotes the estimated indicator in area i based on any of the five methods under consideration and I_i denotes the corresponding true value in area i.

Table 7 presents the results split by the four scenarios. It shows median and mean values of the RMSE and bias averaged over small areas. Under the Normal scenario the EBP without transformation is the gold standard, but the EBP with data-driven transformations (log-shift, Box–Cox and dual power) perform similarly in terms of RMSE and bias. The same picture

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

Table 7. Summaries of estimated RMSEs and bias over the model-based settings

Measure	*Transformation*	*HCR*		*PGAP*		*QSR*	
		Median	*Mean*	*Median*	*Mean*	*Median*	*Mean*
Normal distribution							
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 distribution							
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 distribution							
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 distribution							
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

emerges in the Log-scale scenario where the EBP with a logarithmic transformation is the gold standard, but again the EBP with data-driven transformations perform well in terms of both RMSE and bias. These results confirm our expectations that the EBP with data-driven transformations adapt to the shape of the data distribution. Under the GB2 and Pareto scenarios we note that the EBP with a fixed transformation or without transformation is inferior to the EBP with data-driven transformations in terms of both RMSE and bias. The differences are especially pronounced for QSR which is very sensitive to the tails of the distribution. Further-

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

Table 8. Performance of MSE estimators in model-based simulations: EBP with Box–Cox transformation

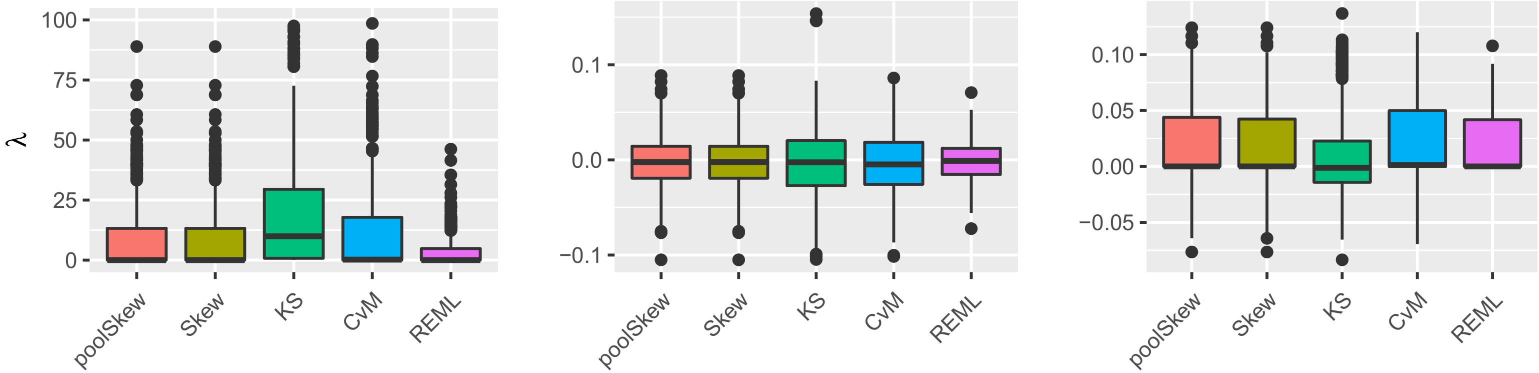
Measure	*Bootstrap*	*HCR*		*PGAP*		*QSR*	
		Median	*Mean*	*Median*	*Mean*	*Median*	*Mean*
Normal distribution							
Relative RMSE (%)	Parametric	8.30	9.22	9.15	9.47	15.25	21.23
	Wild	14.57	14.77	14.21	14.61	17.46	20.93
Relative bias (%)	Parametric	6.64	7.27	−1.17	−0.12	−7.72	−12.61
	Wild	8.05	8.04	2.17	3.23	−1.01	−1.46
Log-scale distribution							
Relative 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
Relative 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 distribution							
Relative 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
Relative 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 distribution							
Relative 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
Relative 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

more, the estimates based on data-driven transformations are almost unbiased or have a small bias. A closer look at the data-driven transformations indicates that the EBP with a log-shift transformation performs slightly better than the EBP with Box–Cox and dual power transformations under the GB2 and Pareto scenarios. Overall, it appears that the proposed EBP method with data-driven transformations adapts to the underlying distribution of the data, and hence improves the precision of small area estimates.

We now turn our attention to the performance of the MSE estimators. We denote by *parametric* and *wild* the proposed parametric bootstrap and proposed semiparametric wild bootstrap respectively. The aim of this part is twofold. Firstly, we assess the performance of the two proposed MSE estimators that we introduced in Section 5. Secondly, we investigate the ability of the wild bootstrap to protect against departures from the assumptions of the unit level error term. Starting with the first aim, Table 8 reports the results for the two MSE estimators and presents the mean and median values of relative RMSE and relative bias—over Monte Carlo replications and areas—of the EBP with Box–Cox transformation. For calculating the RMSE and relative bias we treat the empirical MSE (over Monte Carlo replications) as the true MSE. The results for the EBP with a log-shift transformation and dual power transformation are very similar and although they have been omitted they 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 have small positive relative bias (HCR- and PGAP-indicators) in the Normal and Log-scale settings. However, the parametric bootstrap shows some underestimation in the case of QSR. In this latter case the wild bootstrap appears to be associated with

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025



(a)

(b)

(c)

Fig. 7. Boxplots of estimated transformation parameters for the Log-scale scenario by using different estimation methods: (a) log-shift; (b) Box–Cox; (c) dual

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

Table 9. Mean and median of estimated transformation parameters under the Log-scale scenario by using different estimation methods

Method	*Results for log-shift transformation*		*Results for Box–Cox transformation*		*Results for dual transformation*	
	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

smaller relative bias. For the Normal and Log-scale scenarios the parametric bootstrap also has smaller relative RMSE than the wild bootstrap. Nevertheless, the wild bootstrap provides reasonable results for HCR and PGAP and reduces the underestimation for QSR. When the distributional assumptions are not met, as in the GB2 and Pareto scenarios, the parametric bootstrap clearly tends to underestimate the MSE (except for HCR in the Pareto scenario). Although the wild bootstrap does not completely eliminate this bias, it greatly reduces it and provides more stable MSE estimates in terms of relative RMSE. These results indicate that departures from the model assumptions—even after using data transformations—can impact MSE estimation with parametric methods. The problem is more pronounced when estimating parameters that depend on the tails of the distribution as is the case with QSR. In those cases, the use of the semiparametric bootstrap, at least as a supplementary MSE estimation method, can offer some protection against misspecification.

7.3. Effect of alternative estimation methods for λ

In this last section we explore the use of non-parametric alternatives to the REML approach for estimating data-driven transformation parameters (see Section 4.3). Here, we study five estimation methods. These are the REML approach, the minimization of the skewness *Skew* and the pooled skewness *poolSkew*, and the distance-based criteria Kolmogorov–Smirnov *KS* and Cramér-von Mises *CvM* that we introduced in Section 4.3.

In the scenarios with existing theoretically correct parameters five methods estimate transformation parameters close to them. For instance, in the Log-scale scenario, the estimated transformation parameters under the various estimation methods are shown in Fig. 7 and Table 9. We observe that, although the five methods provide similar estimates of λ, the REML method has smaller variability. In our model-based simulations we further studied the effect of the estimation method of the transformation parameter on point and MSE estimation and we conclude that this only marginally influences the quality of small area estimates. These results are available from the authors on request.

Overall, these results suggest that for the scenarios that we considered in this paper the method that is used to estimate the transformation parameter does not have a noticeable effect on SAE and REML appears to be the most stable method.

8. Conclusions and future research directions

In this paper we investigate data-driven transformations for SAE. In particular, we propose

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

an EBP approach with data-driven transformations estimated with likelihood-based methods. The use of scaled transformations (conditional on the Jacobian) enables the use of standard software for fitting the linear mixed regression model. Three types of transformation were discussed: log-shift, Box–Cox and dual power transformations. We further explore the use of the parametric and semiparametric wild bootstrap for MSE estimation that also captures the uncertainty from estimating the data-driven transformation parameter. The semiparametric bootstrap is used for protecting against departures from the model assumptions. Model-based simulations demonstrate the ability of the proposed EBP method to adapt to the shape of the data distribution and hence provide more efficient estimates than a fixed logarithmic transformation or the case where no transformation is used. Although the paper focuses on the EBP the methods proposed are applicable to other small area estimators, e.g. the approach of Elbers *et al.* (2003). The methods that are proposed in this paper can be implemented by using the R package `emdi` (Kreutzmann *et al.*, 2019). The package supports the user by estimating and mapping regionally disaggregated indicators. Although this package already includes the logarithmic and Box–Cox transformations, some research effort should be shifted towards the development of relevant software which includes in more detail the use of data-driven transformations in the SAE context.

Further research can investigate the use of multiparameter transformation families. This may enable better control of higher moments and hence better adaptation to the distribution of the data. Since likelihood-based approaches might be influenced by outliers, it would be interesting also to investigate robust estimation methods. Model selection with data-driven transformations presents additional challenges. Finding a good working model depends on the method of transformation. In this paper we first find a working model and keep this fixed when considering different data-driven transformations. However, this may not offer the best approach to model selection. Approaches that simultaneously consider both steps for linear regression models have been proposed (Laud and Ibrahim, 1995; Hoeting and Ibrahim, 1998; Hoeting *et al.*, 2002). Extending these approaches to the case of linear mixed models is an open research problem. Finally, comparing the EBP with data-driven transformations to EBP approaches with alternative parametric assumptions (Diallo and Rao, 2014; Graf *et al.*, 2019) is empirical work that remains open.

Acknowledgements

First of all, the authors are indebted to the Joint Editor, Associate Editor and referees for comments that significantly improved the paper. Rojas-Perilla, Schmid and Tzavidis gratefully acknowledge support by grant ES/N011619/1—'Innovations in small area estimation methodologies' from the UK Economic and Social Research Council. The authors are grateful to the Consejo Nacional de Evaluación de la Política de Desarrollo Social for providing the data that were used in the empirical work. The views set out in this paper are those of the authors and do not reflect the official opinion of the Consejo Nacional de Evaluación de la Política de Desarrollo Social. The numerical results are not official estimates and are produced only for illustrating the methods. The authors are grateful for the computation time provided by the High Performance Computing service of the Freie Universität Berlin.

Appendix A: Derivation of scaled transformations

In this appendix we derive the Jacobian and the corresponding scaling factors that were presented in Table 2 for the log-shift, Box–Cox and dual power transformations.

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

A.1. Log-shift transformation

Let $J(\lambda, \mathbf{y})$ be the Jacobian of the log-shift transformation from $\mathbf{y}_i$ to $\mathbf{y}_i^*(\lambda)$, defined as

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

The log-likelihood function in equation (3) can be rewritten as

$$\begin{aligned} l_{\mathrm{REML}}(\mathbf{y}, \lambda|\boldsymbol{\theta}) = &-\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\mathbf{V}_i| - \frac{1}{2}\log\left|\sum_{i=1}^{D}\mathbf{X}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{X}_i\right| \\ &-\frac{1}{2}\sum_{i=1}^{D}(\mathbf{y}_i^*(\lambda) - \mathbf{X}_i\hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{y}_i^*(\lambda) - \mathbf{X}_i\hat{\boldsymbol{\beta}}) - n\log\underbrace{\left\{\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+\lambda)\right\}^{1/n}}_{=\bar{y}_\lambda}. \end{aligned}$$

To obtain the scaled log-shift transformation, $z_{ij}^*(\lambda)$, the denominator of the term

$$y_{ij}^*(\lambda)/J(\lambda, \mathbf{y})^{1/n}$$

is given by

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

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

$$z_{ij}^*(\lambda) = \frac{y_{ij}^*(\lambda)}{J(\lambda, \mathbf{y})^{1/n}} = \bar{y}_\lambda \log(y_{ij} + \lambda)$$

for $y_{ij} > -\lambda$.

A.2. Box–Cox transformation

Let $J(\lambda, \mathbf{y})$ be the Jacobian of the Box–Cox transformation from $\mathbf{y}_i$ to $\mathbf{y}_i^*(\lambda)$, defined as

$$\begin{aligned} J(\lambda, \mathbf{y}) &= \prod_{i=1}^{D} \prod_{j=1}^{n_i} \left| \frac{\mathrm{d}y_{ij}^*(\lambda)}{\mathrm{d}y_{ij}} \right| \\ &= \prod_{i=1}^{D} \prod_{j=1}^{n_i} (y_{ij} + s)^{\lambda-1}. \end{aligned}$$

The log-likelihood function in equation (3) can be rewritten as

$$\begin{aligned} l_{\mathrm{REML}}(\mathbf{y}, \lambda|\boldsymbol{\theta}) = &-\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\mathbf{V}_i| - \frac{1}{2}\log\left|\sum_{i=1}^{D}\mathbf{X}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{X}_i\right| \\ &-\frac{1}{2}\sum_{i=1}^{D}(\mathbf{y}_i^*(\lambda) - \mathbf{X}_i\hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{y}_i^*(\lambda) - \mathbf{X}_i\hat{\boldsymbol{\beta}}) + n(\lambda-1)\log\underbrace{\left\{\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+s)\right\}^{1/n}}_{=\bar{y}}. \end{aligned}$$

To obtain the scaled transformation of the Box–Cox family, $z_{ij}^*(\lambda)$, the denominator of the term $y_{ij}^*(\lambda)/J(\lambda, \mathbf{y})^{1/n}$ is given by

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

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

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

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

for $y_{ij}>-s$.

A.3. Dual power transformation

Let $J(\lambda,\mathbf{y})$ be the Jacobian of the dual power transformation from $\mathbf{y}_i$ to $\mathbf{y}_i^*(\lambda)$, defined as

$$\begin{aligned}J(\lambda,\mathbf{y})&=\prod_{i=1}^{D}\prod_{j=1}^{n_i}\left|\frac{\mathrm{d}y_{ij}^*(\lambda)}{\mathrm{d}y_{ij}}\right|\\&=\prod_{i=1}^{D}\prod_{j=1}^{n_i}\frac{(y_{ij}+s)^{\lambda-1}+(y_{ij}+s)^{-\lambda-1}}{2}.\end{aligned}$$

The log-likelihood function in equation (3) can be rewritten as

$$\begin{aligned}l_{\mathrm{REML}}(\mathbf{y},\lambda|\boldsymbol{\theta})=&-\frac{n-p}{2}\log(2\pi)+\frac{1}{2}\log\left|\sum_{i=1}^{D}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right|-\frac{1}{2}\sum_{i=1}^{D}\log|\mathbf{V}_i|-\frac{1}{2}\log\left|\sum_{i=1}^{D}\mathbf{X}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{X}_i\right|\\&-\frac{1}{2}\sum_{i=1}^{D}(\mathbf{y}_i^*(\lambda)-\mathbf{X}_i\hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{y}_i^*(\lambda)-\mathbf{X}_i\hat{\boldsymbol{\beta}})\\&+n\log\underbrace{\left\{\prod_{i=1}^{D}\prod_{j=1}^{n_i}\frac{(y_{ij}+s)^{\lambda-1}+(y_{ij}+s)^{-\lambda-1}}{2}\right\}^{1/n}}_{=\bar{y}_{\lambda}}.\end{aligned}$$

To obtain the scaled dual transformation $z_{ij}^*(\lambda)$, the denominator of the term $y_{ij}^*(\lambda)/J(\lambda,\mathbf{y})^{1/n}$ is given by

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

Therefore, the scaled dual transformation is defined as follows:

$$z_{ij}^*(\lambda)=\frac{y_{ij}^*(\lambda)}{J(\lambda,\mathbf{y})^{1/n}}=\begin{cases}\bar{y}_{\lambda}^{-1}\dfrac{(y_{ij}+s)^{\lambda}-(y_{ij}+s)^{-\lambda}}{2\lambda} & \text{if } \lambda>0,\\ \bar{y}_{\lambda}^{-1}\log(y_{ij}+s) & \text{if } \lambda=0,\end{cases}$$

for $y_{ij}>-s$.

References

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988) An error component model for prediction of county crop areas using survey and satellite data. *J. Am. Statist. Ass.*, **83**, 28–36.

Bedoya, H., Freije, S., Vila, L., Echeverria, G., Biller, D., Grandolini, G. M., Albisetti, R., Quintrell, E. and Vish, R. (2013) Country partnership strategy for the united Mexican states (2014-2019). *Technical Report*. World Bank Group, Washington DC.

Bickel, P. J. and Doksum, K. A. (1981) An analysis of transformations revisited. *J. Am. Statist. Ass.*, **76**, 296–311.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.

Carroll, R. J. and Ruppert, D. (1987) Diagnostics and robust estimation when transforming the regression model and the response. *Technometrics*, **29**, 287–299.

Chakravarti, I. M., Laha, R. G. and Roy, J. (1967) Handbook of methods of applied statistics. In *Handbook of Methods of Applied Statistics* (eds R. A. Bradley, J. S. Hunter, D. G. Kendall and G. S. Watson). New York: Wiley.

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025

Cramér, H. (1928) On the composition of elementary errors: First paper; Mathematical deductions. *Scand. Act. J.*, 13–74.

Diallo, M. S. and Rao, J. N. K. (2014) Small area estimation of complex parameters under unit-level models with skew-normal errors. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*

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

Elbers, C. and van der Weide, R. (2014) Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality. *Working Paper*. World Bank, Washington DC.

Eurostat (2004) Common cross-sectional EU indicators based on EU-SILC; the gender pay gap. Unit D-2, Directorate D. Eurostat, Luxembourg.

Fabrizi, E. and Trivisano, C. (2016) Small area estimation of the Gini concentration coefficient. *Computnl Statist. Data Anal.*, **99**, 223–234.

Feng, X., He, X. and Hu, J. (2011) Wild bootstrap for quantile regression. *Biometrika*, **98**, 995–999.

Foster, J., Greer, J. and Thorbecke, E. (1984) A class of decomposable poverty measures. *Econometrica*, **52**, 761–766.

González-Manteiga, W., Lombardía, M., Molina, I., Morales, D. and Santamaría, L. (2008) Bootstrap mean squared error of a small-area eblup. *J. Statist. Computn Simuln*, **78**, 443–462.

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

Gurka, M. J., Edwards, L. J., Muller, K. E. and Kupper, L. L. (2006) Extending the Box–Cox transformation to the linear mixed model. *J. R. Statist. Soc.* A, **169**, 273–288.

Hoeting, J. A. and Ibrahim, J. G. (1998) Bayesian predictive simultaneous variable and transformation selection in the linear model. *Computnl Statist. Data Anal.*, **28**, 87–103.

Hoeting, J. A., Raftery, A. E. and Madigan, D. (2002) Bayesian variable and transformation selection in linear regression. *J. Computnl Graph. Statist.*, **3**, 485–507.

John, J. A. and Draper, N. R. (1980) An alternative family of transformations. *Appl. Statist.*, **29**, 190–197.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M. and Tzavidis, N. (2019) emdi: estimating and mapping disaggregated indicators. *J. Statist. Softwr.*, to be published.

Laud, P. W. and Ibrahim, J. G. (1995) Predictive model selection. *J. R. Statist. Soc.* B, **57**, 247–262.

Marino, M. F., Ranalli, M. G., Salvati, N. and Alfo, M. (2019) Semi-parametric empirical best prediction for small area estimation of unemployment indicators. *Ann. Appl. Statist.*, to be published.

Marino, M. F., Tzavidis, N. and Alfo, M. (2018) Mixed hidden Markov quantile regression models for longitudinal data with possibly incomplete sequences. *Statist. Meth. Med. Res.*, **27**, 2231–2246.

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

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

Nakagawa, S. and Schielzeth, H. (2013) A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Meth. Ecol. Evoln*, **4**, 133–142.

R Core Team (2017) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Royston, P. and Lambert, P. C. (2011) *Flexible Parametric Survival Analysis using Stata: beyond the Cox Model*. College Station: StataCorp.

Schmid, T., Bruckschen, F., Salvati, N. and Zbiranski, T. (2017) Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. *J. R. Statist. Soc.* A, **180**, 1163–1190.

Tortajada, C. (2006) Who has access to water case study of Mexico City metropolitan area human development report 2006. *Technical Report*. Third World Center for Water Management, Atizapán.

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 (with discussion). *J. R. Statist. Soc.* A, **181**, 927–979.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*, vol. 1. Berlin: Springer.

Weidenhammer, B., Tzavidis, N., Schmid, T. and Salvati, N. (2014) Domain prediction for counts using microsimulation via quantiles. *Small Area Estimation Conf., Poznan*.

Yang, L. (1995) Transformation-density estimation. *PhD Thesis*. University of North Carolina, Chapel Hill.

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

Yeo, I.-K. and Johnson, R. A. (2000) A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplementary material: Data-driven transformations in small area estimation’.

Downloaded from https://academic.oup.com/jrsssa/article/183/1/121/7056331 by guest on 08 December 2025