Adaptively Transformed Mixed Model Prediction of General Finite Population Parameters

By

Shonosuke Sugasawa and Tatsuya Kubokawa

April 2018

CENTER FOR RESEARCH AND EDUCATION FOR POLICY EVALUATION
DISCUSSION PAPER NO. 19

CENTER FOR RESEARCH AND EDUCATION FOR POLICY EVALUATION (CREPE)
THE UNIVERSITY OF TOKYO
http://www.crepe.e.u-tokyo.ac.jp/

Adaptively Transformed Mixed Model Prediction of General Finite Population Parameters*

SHONOSUKE SUGASAWA

Risk Analysis Research Center, The Institute of Statistical Mathematics

TATSUYA KUBOKAWA

Faculty of Economics, University of Tokyo

Abstract. For estimating area-specific parameters (quantities) in a finite population, a mixed model prediction approach is attractive. However, this approach strongly depends on the normality assumption of the response values although we often encounter a non-normal case in practice. In such a case, transforming observations to make them close to normality is a useful tool, but the problem of selecting suitable transformation still remains open. To overcome the difficulty, we here propose a new empirical best predicting method by using a parametric family of transformations to estimate a suitable transformation based on the data. We suggest a simple estimating method for transformation parameters based on the profile likelihood function, which achieves consistency under some conditions on transformation functions. For measuring variability of point prediction, we construct an empirical Bayes confidence interval of the population parameter of interest. Through simulation studies, we investigate some numerical performances of the proposed methods. Finally, we apply the proposed method to synthetic income data in Spanish provinces in which the resulting estimates indicate that the commonly used log-transformation is not appropriate.

Key words: Confidence interval; Empirical Bayes; Finite population; Mean squared error; Random effect; Small area estimation.

*This version: May 11, 2017

1 Introduction

The mixed model prediction based on random effect models has been widely used in small area estimation (Rao and Molina, 2015). The random effect models used in small area estimation are mainly divided into two models: the Fay-Herriot model (Fay and Herriot, 1979) and the nested error regression model (Battese et al., 1988). Especially, the nested error regression model has been used for estimating population parameters in a finite population. Here we consider a finite population consisting of m areas and each area has N_i units for $i = 1, \ldots, m$. Let Y_{ij} be a characteristic of the jth unit in the ith area, the main purpose is to estimate the area-specific parameter defined as

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} T(Y_{ij}), \tag{1}$$

where $T(\cdot)$ is a known (user-specified) function. The simplest choice is $T(x) = x$, in which μ_i corresponds to the finite population mean, and many literatures have been focused on this case; Chambers et al. (2014), Jiang and Lahiri (2006), Lahiri and Mukherjee (2007) and Schmit et al. (2016). On the other hand, as noted by Molina and Rao (2010), other forms of $T(\cdot)$ are used in practice. For example, in poverty mapping, we often use FGT poverty measure $T(x) = \{(z - x)/z\}^{\alpha} I(x < z)$ (Foster et al., 1984), noting that μ_i corresponds to the poverty rate when $\alpha = 0$.

If all the units Y_{ij} in the ith area were observed, we could calculate the true value of μ_i. However, only a part of the units are available in practice. Let $n_i (< N_i)$ be the number of sampled units and $y_s = \{y_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, m\}$ be the sampled data. It is known that the direct estimator of μ_i using the observed units has high variability, especially in the case that n_i is much smaller than N_i. In real application, some covariates associated with Y_{ij} are available not only for sampled but also for non-sampled units, which are denoted by $\boldsymbol{x}_{ij}$ with $j = 1, \ldots, N_i$ and $i = 1, \ldots, m$. Hence, one aims to estimate μ_i based on the sampled data and information on covariates. To this end, a typical strategy is to assume that all the units follow the nested error

regression model:

$$Y_{ij} = \boldsymbol{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad j = 1, \ldots, N_i, \quad i = 1, \ldots, m, \tag{2}$$

where $\boldsymbol{x}_{ij}$ and $\boldsymbol{\beta}$ are p-dimensional vectors of covariates and regression coefficients, v_i is the area-specific effect which follows $N(0, \tau^2)$ and ε_{ij} is a sampling error distributed as $N(0, \sigma^2)$. Then, the conditional distribution of the non-sampled data Y_{ij} given all the sampled data y_s is given by

$$Y_{ij}|y_s \sim N\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta} + \frac{n_i \tau^2}{\sigma^2 + n_i \tau^2}(\bar{y}_i - \bar{\boldsymbol{x}}_i^t \boldsymbol{\beta}),\ \frac{\sigma^2 \tau^2}{\sigma^2 + n_i \tau^2}\right), \quad j = n_i + 1, \ldots, N_i, \tag{3}$$

which follows from the normality of Y_{ij} under the model (2). Then the best predictor of μ_i under squared error loss is the conditional expectation $\mathrm{E}[\mu_i|y_i]$, which has the form

$$\widetilde{\mu}_i \equiv \mathrm{E}[\mu_i|y_s] = \frac{1}{N_i}\left\{ \sum_{j=1}^{n_i} T(y_{ij}) + \sum_{j=n_i+1}^{N_i} \mathrm{E}[T(Y_{ij})|y_s] \right\}. \tag{4}$$

Here, the expectation $\mathrm{E}[T(Y_{ij})|y_s]$ can be computed via the Monte Carlo integration by generating a large number of random samples from the conditional distribution (3). Moreover, the best predictor $\widetilde{\mu}_i$ depends on the unknown model parameters $\boldsymbol{\beta}, \tau^2$ and σ^2 in the model (2), so that these parameters should be replaced with their estimated counterparts. To this end, one can estimate the model parameters in the model (2) based on the sampled data y_s based on, for example, the maximum likelihood or restricted maximum likelihood methods.

It is observed that the key assumption in deriving the best predictor (4) is the normality of Y_{ij} in (2), which enables us to obtain the simple expression of the conditional distribution (3). However, we often encounter the case where the normality assumption is not plausible for Y_{ij}. In fact, in poverty mapping, Y_{ij} is a welfare variable like income, thereby the distribution of Y_{ij} could be skewed. In this case, Molina and Rao (2010) proposed assuming the nested error model (2) for the transformed variables $H(Y_{ij})$ instead of Y_{ij}. If Y_{ij} is right skewed, one may use $H(x) = \log x$. However, we still suffer from the misspecification of the transformation and the predictor of μ_i

under misspecified transformation would be biased, thereby selecting suitable transformations from the data is desired to validate the prediction method. To overcome the difficulty, we propose the adaptively transformed empirical best predicting method in which we use a parametric family of transformations for data transformation instead of the use of specified transformations. We derive a form of the best predictor of μ_i and provide a simple estimating method for transformation parameters based on profile likelihood function, which produces a consistent estimator under some regularity conditions. We also construct an empirical Bayes confidence interval of μ_i for measuring the variability of the point prediction. The proposed intervals are shown to have $O(m^{-1})$ coverage error, and we also suggest the parametric bootstrap calibration for confidence intervals with further accuracy.

As related methods, Li and Lahiri (2007) suggested to use the Box-Cox transformation (Box and Cox, 1964) for the data transformation for robust estimation of finite population totals while their method was developed under models without random effects. Hence, our method would be more efficient. Concerning the empirical Bayes confidence intervals, Nandram (1999) derived an empirical Bayes confidence intervals of the finite population means, which corresponds to the case taking $T(x) = x$ in (1).

This paper is organized as follows: In Section 2, we describe the proposed prediction method as well as parameter estimation of the model parameters. In Section 3, we construct an empirical Bayes confidence interval of μ_i. In Section 4, we present the results from simulation studies and a data application. In Section 5, we give conclusions and some discussions. The technical proofs are given in Appendix.

2 Adaptively Transformed Empirical Best Prediction

2.1 Transformed best predictor

Let $H_\lambda(\cdot)$ be a family of transformations with parameter λ. The transformation parameter λ might be multidimensional, but we treat λ as a scalar parameter for notational simplicity. The assumptions and specific choices of $H_\lambda(\cdot)$ will be discussed in the subsequent section. We assume that the transformed variable $H_\lambda(y_{ij})$ follows the nested

error regression model:

$$H_\lambda(Y_{ij}) = \boldsymbol{x}_{ij}^t\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad j = 1, \ldots, N_i, \quad i = 1, \ldots, m, \tag{5}$$

where $\boldsymbol{x}_{ij}$ and $\boldsymbol{\beta}$ are p-dimensional vectors of covariates and regression coefficients, v_i and ε_{ij} are an area-specific effect and a sampling error, respectively. Here we assume that v_i and ε_{ij} are mutually independent and distributed as $v_i \sim N(0, \tau^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$ with unknown two variance parameters τ^2 and σ^2. It is worth noting that, owing to the area effect v_i, the units in the same area are mutually correlated while the units in the different area are independent. Specifically, from (5), it holds $\mathrm{Cor}(H_\lambda(Y_{ij}), H_\lambda(Y_{ik})) = (\tau^2 + \sigma^2)^{-1}\tau^2$, $j \neq k$, thereby the units in the same area are mutually correlated and the degree of correlation is determined by the ratio τ^2/σ^2. From the normality assumptions of v_i and ε_{ij}, it follows that $H_\lambda(Y_{ij}) \sim N(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}, \tau^2 + \sigma^2)$. Thus, the transformation parameter λ can be chosen to make the transformed data $H_\lambda(y_{ij})$ close to normality. We define $\boldsymbol{\phi} = (\boldsymbol{\beta}^t, \tau^2, \sigma^2, \lambda)^t$, as the vector of unknown model parameters in (5). The estimation procedure will be given in the subsequent section.

Let $y_s = \{y_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, m\}$ be the sampled data. From the model (5), we have $H_\lambda(Y_{ij})|y_s \sim N(\theta_{ij}, s_i^2 + \sigma^2)$, $j = n_i + 1, \ldots, N_i$, where

$$\theta_{ij} = \boldsymbol{x}_{ij}^t\boldsymbol{\beta} + \frac{\tau^2}{\sigma^2 + n_i\tau^2}\sum_{j=1}^{n_i}(H_\lambda(y_{ij}) - \boldsymbol{x}_{ij}^t\boldsymbol{\beta}), \quad s_i = \sqrt{\frac{\sigma^2\tau^2}{\sigma^2 + n_i\tau^2}}. \tag{6}$$

Hence, the best predictor of μ_i given in (1) can be obtained as

$$\widetilde{\mu}_i(y_s; \boldsymbol{\phi}) \equiv \mathrm{E}[\mu_i|y_s] = \frac{1}{N_i}\left\{\sum_{j=1}^{n_i} T(y_{ij}) + \sum_{j=n_i+1}^{N_i} \mathrm{E}[T \circ H_\lambda^{-1}(u_{ij})]\right\}, \tag{7}$$

where the expectation is taken with respect to $u_{ij} \sim N(\theta_{ij}, s_i^2 + \sigma^2)$, and $T \circ H_\lambda^{-1}(\cdot)$ is the composite function of $T(\cdot)$ and H_λ^{-1}, the inverse function of $H_\lambda(\cdot)$. Although the expectation $\mathrm{E}[T \circ H_\lambda^{-1}(u_{ij})]$ does not have a closed form in general, it can be easily computed via the Monte Carlo integration. We call the best predictor (7) adaptively

transformed best predictor (ATBP).

2.2 Estimation of structural parameters

We here consider estimating the unknown model parameters $\boldsymbol{\phi}$ in (5) based on the marginal likelihood function. Noting that the log-marginal likelihood function of $\boldsymbol{\phi}$ is given by

$$\begin{aligned} L(\boldsymbol{\phi}) = &-\frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{m}\{H_\lambda(y_i) - \boldsymbol{X}_i\boldsymbol{\beta}\}^t\,\boldsymbol{\Sigma}_i^{-1}\,\{H_\lambda(y_i) - \boldsymbol{X}_i\boldsymbol{\beta}\} \\ &- \frac{1}{2}\sum_{i=1}^{m} n_i\log 2\pi + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\log H'_\lambda(y_{ij}), \end{aligned} \tag{8}$$

where $(\boldsymbol{\Sigma}_i)_{k\ell} = \tau^2 + \sigma^2 I(k=\ell)$, $H_\lambda(y_i) = (H_\lambda(y_{i1}), \ldots, H_\lambda(y_{in_i}))^t$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}^t, \ldots, \boldsymbol{x}_{in_i}^t)^t$, and $H'_\lambda(\cdot)$ denotes the derivative of $H_\lambda(\cdot)$. The maximum likelihood estimator of $\boldsymbol{\phi}$ can be defined as the maximizer of $L(\boldsymbol{\phi})$.

For maximizing the likelihood function $L(\boldsymbol{\phi})$, we first note that the profile likelihood function of λ can be expressed as

$$\mathrm{PL}(\lambda) = \mathrm{ML}(\lambda) + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\log H'_\lambda(y_{ij}), \tag{9}$$

where $\mathrm{ML}(\lambda)$ is the maximum likelihood of the nested error regression model with response values $H_\lambda(y_{ij})$ and covariate vectors $\boldsymbol{x}_{ij}$, which can be efficiently carried out by using well-developed numerical method (e.g. Molina and Marhuenda, 2015). Using the ease of the point evaluation of the profile likelihood $\mathrm{PL}(\lambda)$, we can obtain the maximizer of $\mathrm{PL}(\lambda)$ by using, for example, the golden section method (Brent et al., 1973). Once we obtain the estimator $\widehat{\lambda}$, we get the estimators of other parameters by applying the nested error regression model to the data set $\{H_{\widehat{\lambda}}(y_{ij}), \boldsymbol{x}_{ij}\}$.

For estimating the two variance parameters τ^2 and σ^2, the restricted maximum likelihood (RML) method (Jiang, 1996) might be more attractive than the maximum likelihood method. To implement the RML estimation, the first three terms in (8) need to be changed to the restricted maximum likelihood, but the transformation parameter λ can be easily estimated in the same manner as the maximum likelihood

method based on the profile likelihood function. However, in this paper, we consider only the maximum likelihood estimator for simplicity.

2.3 Class of transformations

We here consider the concrete choice of the family of transformations $H_\lambda(\cdot)$. To begin with, we give some conditions to be satisfied by the transformations.

Assumption 1. *(Class of transformations)*

1. *H_λ is a differentiable and monotone function, and the range of H_λ is $\mathbb{R}$ for all λ.*
2. *For fixed x, $H_\lambda(x)$ as the function of λ is differentiable.*
3. *The function $|\partial H_\lambda(w)/\partial\lambda|$, $|\partial^2 H_\lambda(w)/\partial\lambda^2|$ and $|\partial^2 \log H'_\lambda(w)/\partial\lambda^2|$ with $w = H_\lambda^{-1}(x)$ are bounded from the upper by $C_1\{\exp(C_2 x) + \exp(-C_2 x)\}$ with some constants $C_1, C_2 > 0$.*

The first condition is crucial in this context. If the range of H_λ is not $\mathbb{R}$, but some subset $A \subset \mathbb{R}$, the inverse function H_λ^{-1} cannot be defined on $\mathbb{R} \setminus A$, which causes problems in computing the best predictor (7). When the observations are positive valued, the Box-Cox (BC) transformation (Box and Cox, 1964), $H_\lambda(x) = \lambda^{-1}(x^\lambda - 1)$ for $\lambda \neq 0$ and $H_0(x) = \log(x)$, is widely used. However, it is known that the range of BC transformation is truncated and not whole real line, so that the BC transformation cannot be used in this context. An alternative transformation, called dual power (DP) transformation, has been suggested by Yang (2006):

$$H_\lambda^{\mathrm{DP}}(x) = \frac{x^\lambda - x^{-\lambda}}{2\lambda}, \qquad x > 0, \quad \lambda > 0, \tag{10}$$

where $\lim_{\lambda\to 0} H_\lambda^{\mathrm{DP}}(x) = \log x$. It can be seen as the mean of two BC transformations, and it is easy to confirm that the range of DPT is $\mathbb{R}$, so that DPT can be used as a parametric family including log-transformation in this context. The expression of the inverse function is required in computing the transformed best predictor (7), and the

Jacobian is also needed for computing the profile likelihood function (9). These are given by

$$H_\lambda^{\mathrm{DP}(-1)}(x) = \left(\lambda x + \sqrt{1+\lambda^2 x^2}\right)^{1/\lambda} \quad \text{and} \quad \frac{dH_\lambda^{\mathrm{DP}}(x)}{dx} = \frac{1}{2}(x^{\lambda-1} + x^{-\lambda-1}).$$

In the context of small area estimation, the DP transformation was used in Sugasawa and Kubokawa (2017) in the Fay-Herriot model. The original DP transformation (10) can be used when the response variables are positive. When response variables are real valued, one may use the shifted-DP transformation of the form $H_{\lambda,c}(x) = \{(x+c)^\lambda - (x+c)^{-\lambda}\}/2\lambda$, where $c \in (\min(y_{ij}) + \varepsilon, \infty)$ with specified small $\varepsilon > 0$.

Another attractive transformation is the sinh-arcsinh (SS) transformation suggested in Jones and Pewsey (2009) in the context of distribution theory, which has the form

$$H_{a,b}^{\mathrm{SS}}(x) = \sinh(b\sinh^{-1}(x) - a), \quad x \in (-\infty, \infty), \quad a \in (-\infty, \infty), \quad b \in (0, \infty) \quad (11)$$

where $\sinh(x) = (e^x - e^{-x})/2$ is the hyperbolic sine function, $\sinh^{-1}(x) = \log(x + \sqrt{x^2+1})$, and two transformation parameter a and b control skewness and tail heaviness, respectively. The inverse transformation and the Jacobian are obtained as

$$H_{a,b}^{\mathrm{SS}(-1)}(x) = \sinh(b^{-1}\sinh^{-1}(x) + a), \quad \text{and} \quad \frac{dH_{a,b}^{\mathrm{SS}}(x)}{dx} = b\sqrt{\frac{1+H_{a,b}^{\mathrm{SS}}(x)^2}{1+x^2}}.$$

These transformations will be used and compared in the application presented in Section 4.3.

2.4 Large sample properties

We here consider the large sample properties of the estimator of structural parameters. To this end, we assume the following condition:

Assumption 2. *(Assumptions under large m)*

1. *The true parameter vector $\boldsymbol{\phi}_0$ is an interior point of the parameter space $\boldsymbol{\Phi}$.*

2. $0 < \min_{i=1,\ldots,m} N_i \leq \max_{i=1,\ldots,m} N_i < \infty$.

3. The elements of $\boldsymbol{X}_i$ *are uniformly bounded and* $\boldsymbol{X}_i^t \boldsymbol{X}_i$ *is positive definite.*

4. $m^{-1} \sum_{i=1}^m \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i$ *converges to a positive definite matrix as* $m \to \infty$.

Since the asymptotic variance and covariance matrix of MLE can be derived from the Fisher information matrix, we first provide the Fisher information matrix in the following Theorem, where the proof is given in Appendix.

Theorem 1. *We define the Fisher information* $I_{\phi_k \phi_j} = -\mathrm{E}[\partial^2 L(\boldsymbol{\phi}) / \partial \phi_k \partial \phi_j]$*, then it follows that*

$$
\begin{aligned}
&I_{\tau^2\tau^2} = \frac{1}{2} \sum_{i=1}^m (\mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i})^2, \quad I_{\tau^2\sigma^2} = \frac{1}{2} \sum_{i=1}^m \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-2} \mathbf{1}_{n_i}, \quad I_{\sigma^2\sigma^2} = \frac{1}{2} \sum_{i=1}^m \mathrm{tr}\,(\boldsymbol{\Sigma}_i^{-2}), \\
&\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sum_{i=1}^m \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i, \quad \boldsymbol{I}_{\boldsymbol{\beta}\tau^2} = \boldsymbol{I}_{\boldsymbol{\beta}\sigma^2} = 0, \quad I_{\lambda\sigma^2} = -\sum_{i=1}^m \mathrm{E}\left[\boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-2} H_\lambda^{(1)}(y_i) \right], \\
&\boldsymbol{I}_{\lambda\boldsymbol{\beta}} = -\sum_{i=1}^m \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \mathrm{E}\left[H_\lambda^{(1)}(y_i) \right], \quad I_{\lambda\tau^2} = -\sum_{i=1}^m \mathrm{E}\left[\boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1} H_\lambda^{(1)}(y_i) \right], \\
&I_{\lambda\lambda} = \sum_{i=1}^m \mathrm{E}\left[H_\lambda^{(1)}(y_i)^t \boldsymbol{\Sigma}_i^{-1} H_\lambda^{(1)}(y_i) \right] + \sum_{i=1}^m \mathrm{E}\left[\boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-1} H_\lambda^{(2)}(y_i) \right] - \sum_{i=1}^m \sum_{j=1}^{n_i} \mathrm{E}\left[\frac{\partial^2}{\partial \lambda^2} \log H_\lambda'(y_{ij}) \right],
\end{aligned}
$$

where $H_\lambda^{(k)}(y_i) = \partial^k H_\lambda(y_i) / \partial \lambda^k$ *for* $k = 1, 2$, $\boldsymbol{z}_i = H_\lambda(y_i) - \boldsymbol{X}_i \boldsymbol{\beta}$, *and* $\mathrm{E}[\cdot]$ *denotes the expectation with respect to* y_{ij}*'s following the model (5). Then, under Assumptions 1 and 2, the maximum likelihood estimator* $\widehat{\boldsymbol{\phi}}$ *is asymptotically distributed as* $\widehat{\boldsymbol{\phi}} \sim N(\boldsymbol{\phi}, \boldsymbol{I}_{\boldsymbol{\phi}}^{-1})$.

From Theorem 1, it is observed that the information matrix of $(\boldsymbol{\beta}^t, \tau^2, \sigma^2)$ does not depend on the transformation parameter λ, and their expressions are the same as those of the traditional nested error regression models. While the two variance parameters τ^2 and σ^2 are orthogonal to $\boldsymbol{\beta}$ in the sense that $\boldsymbol{I}_{\boldsymbol{\beta}\tau^2} = \boldsymbol{I}_{\boldsymbol{\beta}\sigma^2} = 0$, the transformation parameter λ is not orthogonal to the others. The expectations appeared in the Fisher matrix is not analytically tractable, but it can be easily estimated by replacing the expectation with its sample counterpart. In the case that λ is multidimensional, the extension of Theorem 1 is straightforward. The expressions of $H_\lambda^{(k)}(y_i)$ and $\partial^2 \log H_\lambda'(y_{ij}) / \partial \lambda^2$ could be analytically complicated and require tedious algebraic

calculations. In such a case, the numerical derivative can be useful since we need to compute only the point values of the derivatives.

3 Empirical Bayes Confidence Intervals

3.1 Asymptotically valid confidence intervals

Measuring the variability of the transformed empirical best predictor $\widehat{\mu}_i$ is an important issue in practice. Traditionally, the mean squared error (MSE) of $\widehat{\mu}_i$ has been used, and several methods ranging from analytical method (Prasad and Rao, 1990) to numerical methods (Hall and Maiti, 2006) have been considered. On the other hand, an empirical Bayes confidence interval of μ_i is more preferable since it can provide distributional information than MSE though construction of the confidence interval is generally difficult. Here, we derive an asymptotically valid empirical Bayes confidence interval of μ_i.

The key to the confidence interval is the conditional distribution of μ_i given y_s. Noting that $\mathrm{Cov}(H_\lambda(Y_{ij}), H_\lambda(Y_{ik})|y_s) = \mathrm{Var}(v_i|y_s) = s_i^2$ for $j \neq k$, it follows that

$$(H_\lambda(Y_{i,n_i+1}), \ldots, H_\lambda(Y_{iN_i}))^t | y_s \sim N((\theta_{i,n_i+1}, \ldots, \theta_{iN_i})^t, s_i^2 \mathbf{1}_{N_i-n_i}\mathbf{1}_{N_i-n_i}^t + \sigma^2 \boldsymbol{I}_{N_i-n_i}),$$

namely, the each component has the expression

$$H_\lambda(Y_{ij})|y_i = \theta_{ij} + s_i z_i + \sigma w_{ij}, \quad j = n_i+1, \ldots, N_i,$$

where z_i and w_{ij} are mutually independent standard normal random variables, and θ_{ij} and s_i are defined in (6). Then the posterior distribution of μ_i can be expressed as

$$\mu_i|y_i \stackrel{\mathrm{d}}{=} \frac{1}{N_i}\left\{\sum_{j=1}^{n_i} T(y_{ij}) + \sum_{j=n_i+1}^{N_i} T \circ H_\lambda^{-1}\left(\theta_{ij} + s_i z_i + \sigma w_{ij}\right)\right\}, \tag{12}$$

which is a complex function of standard normal random variables z_i and w_{ij}. However, random samples from the conditional distribution (12) can be easily simulated.

We define $Q_a(y_i, \boldsymbol{\phi})$ as the lower $100a\%$ quantile point of the posterior distri-

bution of μ_i with the true $\boldsymbol{\phi}$, which satisfies $\mathrm{P}(\mu_i \leq Q_a(y_i, \boldsymbol{\phi})|y_i) = a$. Hence, the Bayes confidence interval of μ_i with nominal level $1-\alpha$ is obtained as $I_\alpha = (Q_{\alpha/2}(y_i, \boldsymbol{\phi}), Q_{1-\alpha/2}(y_i, \boldsymbol{\phi}))$, which holds that $\mathrm{P}(\mu_i \in I_\alpha) = 1-\alpha$. However, the interval I_α depends on the unknown parameter $\boldsymbol{\phi}$, so that the feasible version of I_α is obtained by replacing $\boldsymbol{\phi}$ with its estimator $\widehat{\boldsymbol{\phi}}$, namely

$$I_\alpha^N = (Q_{\alpha/2}(y_s, \widehat{\boldsymbol{\phi}}), Q_{1-\alpha/2}(y_s, \widehat{\boldsymbol{\phi}})), \tag{13}$$

which we call naive empirical Bayes confidence interval of μ_i. The two quantiles appeared in (13) can be computed by generating a large number of random samples from the conditional distribution (12). Owing to the asymptotic properties of $\widehat{\boldsymbol{\phi}}$, the coverage probability of the naive interval (13) converges to the nominal level as the number of areas m tends to infinity as shown in the following theorem proved in Appendix.

Theorem 2. *Under Assumptions 1 and 2, it holds* $\mathrm{P}(\mu_i \in I_\alpha^N) = 1-\alpha+O(m^{-1})$.

3.2 Bootstrap calibrated intervals

As shown in Theorem 2, the coverage error of the naive interval (13) is of order m^{-1}, which is not necessarily negligible when m is not sufficiently large. Since the number of m is usually moderate in practice, the calibrated intervals with higher accuracy would be valuable. Following Chatterjee, et al. (2008), Hall and Maiti (2006), we construct a second order corrected empirical Bayes confidence interval I_α^C satisfying $P(\mu_i \in I_\alpha^C) = 1-\alpha+o(m^{-1})$.

To begin with, we define the bootstrap estimator of the coverage probability of the naive interval. Let Y_{ij}^* be the parametric bootstrap samples generated from the estimated model (5) with $\boldsymbol{\phi} = \widehat{\boldsymbol{\phi}}$, and $y_s^* = \{Y_{ij}^*,\ j = 1, \ldots, n_i,\ i = 1, \ldots, m\}$. Moreover, let μ_i^* be the bootstrap version of μ_i based on Y_{ij}^*'s. Since the coverage probability is $\mathrm{P}(Q_{a/2}(y_s, \widehat{\boldsymbol{\phi}}) \leq \mu_i \leq Q_{1-a/2}(y_s, \widehat{\boldsymbol{\phi}}))$, its parametric bootstrap estimator can be defined as

$$\mathrm{CP}(a) = \mathrm{E}^* \left[I \left\{ Q_{a/2}(y_s^*, \widehat{\boldsymbol{\phi}}) \leq \mu_i^* \leq Q_{1-a/2}(y_s^*, \widehat{\boldsymbol{\phi}}) \right\} \right],$$

where the expectation is taken with respect to the bootstrap samples Y_{ij}^*'s. Based on the coverage probability, we define the calibrated nominal level a^* as the solution of the equation $\mathrm{CP}(a^*) = 1-\alpha$, which can be solved by the bisectional method (Brent, 1973). Then, the calibrated interval is given by

$$I_\alpha^C = (Q_{a^*/2}(y_s, \widehat{\phi}), Q_{1-a^*/2}(y_s, \widehat{\phi})), \tag{14}$$

which has second order accuracy as shown in the following theorem proved in Appendix.

Theorem 3. *Under Assumptions 1 and 2, it holds* $\mathrm{P}(\mu_i \in I_\alpha^C) = 1-\alpha+o(m^{-1})$.

4 Numerical Studies

4.1 Evaluation of prediction errors

We first evaluate the prediction errors of the proposed predictors together with some existing methods. To this end, we considered the following data generating processes:

(A) $(2\lambda)^{-1}(Y_{ij}^\lambda - Y_{ij}^{-\lambda}) = \beta_0 + \beta_1 X_{ij} + v_i + \varepsilon_{ij}, \quad v_i \sim N(0,\tau^2), \quad \varepsilon_{ij} \sim N(0,\sigma^2)$

(B) $(2\lambda)^{-1}(Y_{ij}^\lambda - Y_{ij}^{-\lambda}) = \beta_0 + \beta_1 X_{ij} + v_i + \varepsilon_{ij}, \quad v_i \sim t_5(0,\tau^2), \quad \varepsilon_{ij} \sim t_5(0,\sigma^2)$

(C) $Y_{ij} = \exp(\beta_0 + \beta_1 X_{ij}) v_i \varepsilon_{ij}, \quad v_i \sim \Gamma(1/\tau^2, 1/\tau^2), \quad \varepsilon_{ij} \sim \Gamma(1/\sigma^2, 1/\sigma^2)$

(D) $Y_{ij} = 0.2\exp(U_{ij}) + 0.8U_{ij}^2, \quad U_{ij} = \beta_0 + \beta_1 X_{ij} + v_i + \varepsilon_{ij},$
$v_i \sim N(0,\tau^2), \quad \varepsilon_{ij} \sim N(0,\sigma^2),$

where $i = 1,\ldots,m$, $j = 1,\ldots,N$, $\beta_0 = -1$, $\beta_1 = 3$, $\tau = 0.3$, $\sigma = 0.7$, and X_{ij} were initially generated from $U(1,2)$ and fixed through simulation experiments. In model (i) and (ii), we considered three values for λ, $\lambda = 0, 0.2$ and 0.4. In this study, we set $N = 200$ and $m = 25$, and we focus on estimating the ratio of the observation with values under z, namely

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} I(Y_{ij} < z), \qquad i = 1,\ldots,m, \tag{15}$$

where z is defined as 0.6 times median of Y_{ij}'s.

Concerning the area sample sizes, we divided $m = 25$ areas into five groups with equal number of areas, and we set the same number of n_i within the same groups. The group pattern of n_i we considered was $(20, 40, 60, 80, 100)$. Among the generated $Y_{i1}, \ldots, Y_{iN}$, we used first n_i observations $y_{i1}(= Y_{i1}), \ldots, y_{in_i}(= Y_{in_i})$ as the sampled data. Then, based on the sampled data y_{ij}'s and covariates X_{ij}'s, we computed the predicted value of μ_i based on the four methods: the proposed flexible transformed empirical best prediction (ATP) method with DP transformation (10), the transformed empirical best prediction (TP) method proposed by Molina and Rao (2010) with log-transformation, the empirical best prediction (EBP) method by directly applying the nested error regression model to the non-transfdemd observation y_{ij}, and the direct estimator (DE) given by

$$\widehat{\mu}_i^D = \frac{1}{n_i} \sum_{j=1}^{n_i} I(y_{ij} < z), \qquad i = 1, \ldots, m.$$

It should be noted that the TP method is correctly specified in scenario (A) with $\lambda = 0$ while the ATP method is overfitting in this case. In the other cases in scenario (A), the ATP method uses the same model as the data generating model. Scenario (B) is similar to (A), but the distribution of error terms have the t-distribution. In scenario (C) and (D), the data generation models do not coincides with any methods.

To compare the performances of the four methods, we computed the square root of mean squared error (RMSE) defined as

$$\text{RMSE}_i = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left(\widehat{\mu}_i^{(r)} - \mu_i^{(r)} \right)^2},$$

where $R = 2000$ in this study, $\widehat{\mu}_i^{(r)}$ and $\mu_i^{(r)}$ are the estimated and true values of μ_i, respectively, in the rth iteration. The obtained values of RMSEs are averaged over the same groups and the results are reported in Table 1.

From Table 1, we can observe that the proposed method provides better estimates than three existing methods in almost all cases. As mentioned in the above, ATP method is overfitting in scenario (A) with $\lambda = 0$ while TP method is correctly specified.

However, the results show that the performances between ATP and TP are almost the same, which might indicate that the MSE inflation due to overfilling is not serious. The similar observation can be done in scenario (B) with $\lambda = 0$. On the other hand, in the other cases, the proposed ATP method can improve the estimation accuracy of TP method as well as EBP and DE methods, by adaptively estimating the transformation parameter from the data.

4.2 Finite sample evaluation of empirical Bayes confidence intervals

We next evaluate the finite sample performances of the empirical Bayes confidence intervals given in Section 3. To this end, we considered the following data generating process for population variables Y_{ij}:

$$(2\lambda)^{-1}(Y_{ij}^{\lambda} - Y_{ij}^{-\lambda}) = \beta_0 + \beta_1 X_{ij} + v_i + \varepsilon_{ij}, \quad v_i \sim N(0, \tau^2), \quad \varepsilon_{ij} \sim N(0, \sigma^2),$$

where $j = 1, \ldots, N$ and $i = 1, \ldots, m$ with $n = 200$. We set the true parameter values $\lambda = 0.3$, $\beta_0 = -1$, $\beta_1 = 3$, $\tau = 0.3$, $\sigma = 0.7$, and X_{ij} were initially generated from the uniform distribution on $(1, 2)$, which were fixed through simulation runs. We focused on the same population parameter given in (15).

Among the generated $Y_{i1}, \ldots, Y_{iN}$, the first $n = 50$ observations $Y_{i1}, \ldots, Y_{in}$ were used as the sampled data $y_{i1}, \ldots, y_{in}$. Then, based on y_{ij}'s and X_{ij}'s, we computed two types pf confidence intervals for μ_i, naive confidence interval (13) and bootstrap calibrated confidence interval (14), which are denoted by NCI and BCI, respectively. To evaluate the performances of two confidence intervals, based on $R = 2000$ simulation runs, we computed the empirical coverage probability (CP) and the average length of confidence interval (AL), which are defined as

$$\mathrm{CP}_i = \frac{1}{R}\sum_{r=1}^{R} I(\mu_i^{(r)} \in \mathrm{CI}_i^{(r)}) \quad \text{and} \quad \mathrm{AL}_i = \frac{1}{R}\sum_{r=1}^{R} |\mathrm{CI}_i^{(r)}|,$$

where $\mu_i^{(r)}$ is the true value and $\mathrm{CI}_i^{(r)}$ is NCI or BCI in the rth iteration. In Figure, we show the obtained CP and AL in each area for two cases $m = 20$ and $m = 30$.

Table 1: The group-wise averaged values of simulated square root of mean squared errors (RMSE) for four methods, proposed adaptively transformed prediction (ATP) method, Molina and Rao's transformed prediction (TP) method, empirical best prediction (EBP) method without any data transformations, and direct estimator (DE) for eight scenarios. All the values in the table are multiplied by 100.

		Area sample size n_i				
Scenario	Method	20	40	60	80	100
(A) $\lambda = 0$	ATP	4.33	3.27	2.86	2.40	2.04
	TP	4.33	3.27	2.85	2.40	2.04
	EBP	6.20	4.73	4.18	3.61	2.80
	DE	8.93	5.78	4.68	3.93	3.80
(A) $\lambda = 0.2$	ATP	4.28	3.25	2.80	2.39	1.90
	TP	4.42	3.38	2.94	2.54	1.99
	EBP	6.24	4.64	4.09	3.68	2.58
	DE	8.34	5.67	4.65	3.80	3.63
(A) $\lambda = 0.4$	ATP	4.06	3.05	2.65	2.29	1.83
	TP	4.61	3.51	3.16	2.82	2.16
	EBP	4.84	3.58	3.08	2.74	2.07
	DE	8.45	5.34	4.48	3.43	3.33
(B) $\lambda = 0$	ATP	4.58	3.33	2.81	2.31	2.01
	TP	4.58	3.33	2.80	2.31	2.01
	EBP	8.92	7.38	6.10	5.76	4.47
	DE	8.74	6.37	5.79	3.75	2.60
(B) $\lambda = 0.2$	ATP	4.33	3.42	2.85	2.28	1.95
	TP	4.56	3.61	3.03	2.45	2.09
	EBP	6.23	5.08	4.52	3.50	3.01
	DE	8.24	6.45	5.73	3.64	2.58
(B) $\lambda = 0.4$	ATP	4.13	3.25	2.70	2.19	1.93
	TP	4.73	3.88	3.23	2.71	2.30
	EBP	4.68	3.72	3.22	2.61	2.26
	DE	7.82	5.93	5.33	3.46	2.56
(C)	ATP	4.90	3.63	2.96	2.41	2.17
	TP	5.02	3.69	3.03	2.47	2.20
	EBP	6.78	5.74	4.36	3.27	3.11
	DE	8.67	5.31	4.16	4.07	3.05
(D)	ATP	4.54	3.44	2.98	2.53	2.03
	TP	5.05	4.04	3.48	2.97	2.36
	EBP	5.25	4.20	3.38	2.90	2.32
	DE	9.85	5.76	4.74	3.60	3.45

Concerning CP, the naive method tends to produce shorter confidence intervals, so that the coverage probability is smaller than the nominal level for all areas, which is more serious in case $m = 20$ than $m = 30$. This comes from the accuracy of NCI presented in Theorem 2, which mentions that the coverage accuracy of NCI is $O(m^{-1})$. On the

other hand, bootstrap method can improve the drawbacks of the naive method, and provides reasonable CP around the nominal level under both $m = 20$ and $m = 30$. The results clearly support the theoretical property given in Theorem 3 presenting BCI is second order accurate. Since undervaluation of estimation risk may produce serious problems in practice, we should be use the bootstrap method when the number of areas is not large.

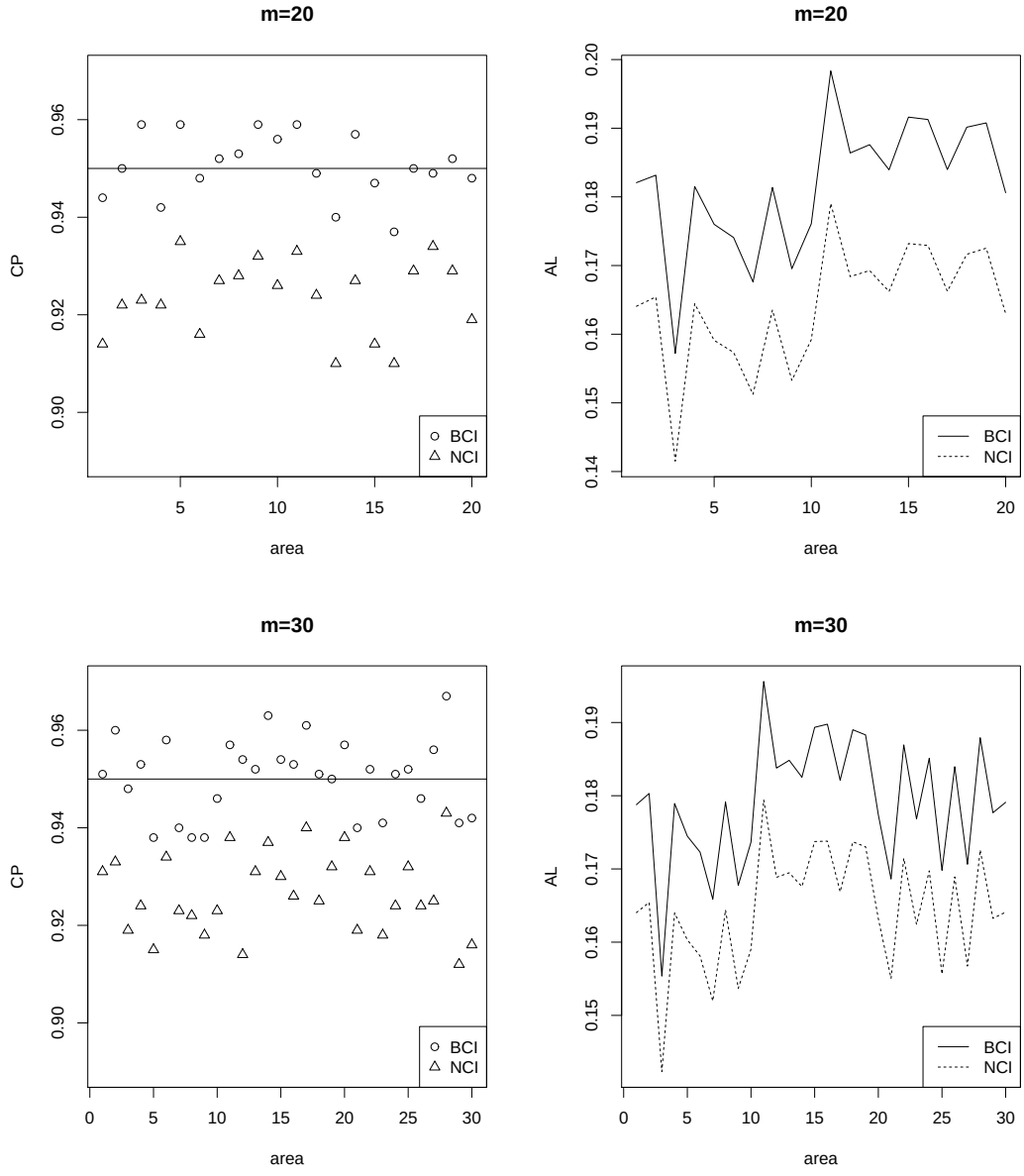

Figure 1: Simulated coverage probability (CP) and average length (AL) of two confidence intervals, naive confidence interval (NCI) and bootstrap calibrated confidence interval (BCI) for $m = 20$ (upper) and $m = 30$ (lower).

4.3 Example: poverty mapping in Spain

We applied the proposed method to estimation of poverty indicators in Spanish provinces, using the synthetic income data available in `sae` package (Molina and Marhuenda, 2015) in `R` language, in which the equalized annual net income are given. The similar data set was used in Molina and Rao (2010) and Molina et al. (2014). As auxiliary variables, we considered the indicators of the five quinquennial groupings of the variable age, the indicator of having Spanish nationality, the indicators of the three levels of the variable education level, and the indicators of the three categories of the variable employment, with categories unemployed, employed and inactive. For each auxiliary variable, one of the categories was considered as base reference, omitting the corresponding indicator and then including an intercept in the model. The poverty measures we focused on were the FGT poverty measures (Foster et al., 1984):

$$T(x) = \left(\frac{x-z}{z}\right)^{\alpha} I(x < z),$$

where z is a fixed poverty line, and it corresponds to poverty incidence or head count ratio ($\alpha = 0$), poverty gap ($\alpha = 1$) and poverty severalty ($\alpha = 2$). In this example, we focused on poverty ratio ($\alpha = 0$), and we set z as the 0.6 times the median of incomes. Let E_{ij} be the income of jth individual in ith area. Such data are available for $m = 52$ areas and the sample sizes are are ranging from 20 to 1420. Since the small portion of E_{ij} take negative values, we assume the nested error regression model with shifted-DPT:

$$\text{SDP:} \quad (2\lambda)^{-1}\left\{(E_{ij}+c)^{1/\lambda} - (E_{ij}+c)^{-1/\lambda}\right\} = \boldsymbol{x}_{ij}^t\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \tag{16}$$

noting that the model has two transformation parameters λ and c. We also considered two submodel of (16). In both models, we set $c = c^* \equiv \min(E_{ij}) + 1$ to ensure that $E_{ij} + c^*$ is positive for all (i, j). The first submodel is denied by putting $c = c^*$ in (16), which is referred to SDP-s. The second sub-model is the shifted-log transformation

model:

$$\text{SL:} \quad \log(E_{ij} + c^*) = \boldsymbol{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \tag{17}$$

which has no longer parameters and was used in Molina and Rao (2010). Finally, we also applied the model with sinh-arcsinh transformation presented in Section 2.3:

$$\text{SS:} \quad \sinh(b \sinh^{-1}(E_{ij}) - a) = \boldsymbol{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \tag{18}$$

which has two transformation parameter a and b.

By maximizing the profile likelihood function of transformation parameters, we obtained as follows:

$$\begin{aligned}
&\text{(SDP)} \ \ \widehat{\lambda} = 0.090 \ (1.99 \times 10^{-3}), \quad \widehat{c} = 4319 \ (170.69) \\
&\text{(SDP-s)} \ \ \widehat{\lambda} = 0.290 \ (8.18 \times 10^{-4}) \\
&\text{(SS)} \ \widehat{a} = -0.584 \ (8.06 \times 10^{-4}), \quad \widehat{b} = 0.463 \ (1.55 \times 10^{-6}),
\end{aligned}$$

where the values in the parentheses are the corresponding standard errors calculated from the Fisher information matrix given in Theorem 1. From the above result, it can be observed that the approximate 95% confidence intervals of the transformation parameter λ in SDP and SDP-s are bounded from 0, which means that the log-transformed model would be inappropriate. Moreover, we computed AIC and BIC based on the maximum marginal likelihood, and the results are given in Table 2 in which the values scaled by the number of sampled units ($N = 17199$) are reported. The results show that the SDP fits the best among the four models in terms of both AIC and BIC while the SL model fits the worst. Hence, the use of parametric transformation can improve AIC and BIC in this application. To see the fitting of the models in terms of normality assumption of the error terms, we computed the standardized residuals defined as

$$r_{ij} = \frac{\widehat{H}(y_{ij}) - \boldsymbol{x}_{ij}^t \widehat{\boldsymbol{\beta}}}{\sqrt{\widehat{\tau}^2 + \widehat{\sigma}^2}}, \quad j = 1, \ldots, n_i, \ \ i = 1, \ldots, m,$$

where $\widehat{H}$ is the estimated transformation function, noting that r_{ij}'s asymptotically follow the standard normal distributions if the assumed model is correctly specified. In Figure 2, we shows QQ-plots of r_{ij}'s of the four models. We can observe that the normality assumptions in the three models with parametric transformations, SDP, SDP-s and SS, seem plausible from Figure 2. However, the QQ-plot for SL shows that the distribution of standardized residuals is skewed and the normality assumption would not be appropriate.

Finally, we calculated the estimated values of the poverty rates μ_i from the direct estimator (DE), and four model based methods. For computing the empirical best predictor of μ_i, we used 500 random samples for Monte Carlo integration. The obtained values are given in Table 3. In the parentheses, we provided the bootstrap empirical Bayes confidence intervals of μ_i with 200 bootstrap iterations. It can be seen that the direct estimator produces quite different estimates of μ_i from the model based methods when area sample size n_i is not large like in Avila and Tarragona. On the other hand, in provinces with large samples sizes, the differences of estimates between DE and the other model methods are relatively small. We can also observe that SL method tend to produce smaller estimates than the other model based methods. However, from AIC and BIC values and QQ-plot in Figure 2, the validity of SL method is highly doubtful in this case, so that the predicted values given in Table 3 would not be reliable. As shown in Table 3, the use of different transformation function leads to significantly different predicted values of μ_i. Hence, it would be valuable to select an adequate transformation function by estimating transformation parameters based on the sampled data.

Table 2: AIC and BIC of four models. The values are scaled by the number of sampled units ($N = 17199$).

	SDP	SDP-s	SS	SL
AIC	20.2241	20.2260	20.2415	20.2883
BIC	20.2305	20.2318	20.2478	20.2937

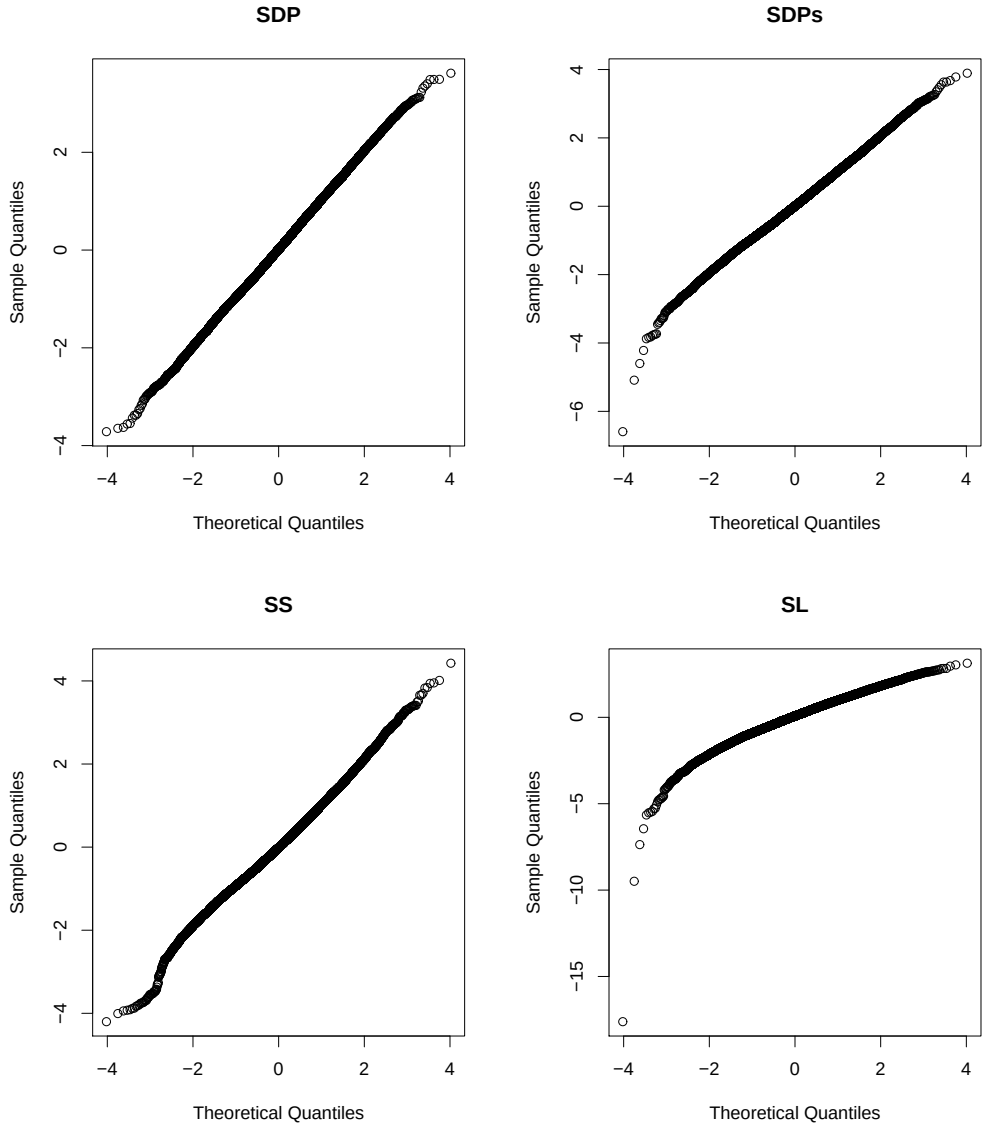

Figure 2: QQ-plots of standardized residuals in four models.

5 Conclusions and Discussion

We have introduced the use of a parametric family of transformations for estimating (predicting) general area specific parameters based on the mixed effects models. We have provided the best predictor of the parameter as well as the maximum likelihood method for estimating model parameters. Moreover, for measuring variability of the predictor, we constructed an empirical Bayes confidence interval of the area parameter. The simulation and empirical studies have revealed that the use of parametric

Table 3: Estimated poverty rates from the direct estimator (DE) and four model based methods in five provinces. The bootstrap empirical Bayes confidence intervals are given in the parenthesis.

area	n_i	DE	SDP	SDP-s	SS	SL
Avila	58	0.57	2.71	2.84	3.12	2.42
			(1.63, 4.05)	(1.87, 4.31)	(2.00, 4.50)	(1.57, 3.62)
Tarragona	134	9.43	8.19	8.48	9.07	7.44
			(6.02, 10.42)	(6.38, 10.82)	(7.28 ,11.27)	(5.83, 9.35)
Santander	434	6.16	6.79	7.02	7.39	6.09
			(5.75, 8.07)	(6.03, 8.03)	(6.38, 8.49)	(5.21, 6.99)
Sevilla	482	3.70	4.95	5.18	5.53	4.29
			(4.15, 5.75)	(4.45 ,6.07)	(4.77, 6.35)	(3.73, 4.90)
Oviedo	803	5.24	4.78	4.97	5.36	4.25
			(4.14, 5.44)	(4.45, 5.51)	(4.75, 6.00)	(3.77, 4.74)

transformations would improve the prediction accuracy of the existing method using specified transformations.

Although we considered an empirical Bayes approach in this paper, the hierarchical Bayes approach as considered in Molina et al. (2014), by assigning some prior distributions for model parameters, would be useful. Moreover, one may use more flexible method for the regression part like the penalized spline as used in Opsomer et al. (2008). The detailed investigation of these issues are left to a valuable future study.

Acknowledgement Shonosuke Sugasawa and Tatsuya Kubokawa are supported by Grant-in-Aid for Scientific Research (16H07406, 15H01943 and 26330036) from Japan Society for the Promotion of Science.

Appendix

A1. Proof of Theorem 1. From the likelihood function (8), its first order deriva-

tives are given by

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m} \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{z}_i, \quad \frac{\partial L}{\partial \tau^2} = -\frac{1}{2} \sum_{i=1}^{m} \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} - \frac{1}{2} \sum_{i=1}^{m} \boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{z}_i$$
$$\frac{\partial L}{\partial \sigma^2} = -\frac{1}{2} \sum_{i=1}^{m} \mathrm{tr}\,(\boldsymbol{\Sigma}_i^{-1}) - \frac{1}{2} \sum_{i=1}^{m} \boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-2} \boldsymbol{z}_i,$$
$$\frac{\partial L}{\partial \lambda} = -\sum_{i=1}^{m} \boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-1} H_\lambda^{(1)}(y_i) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{\partial}{\partial \lambda} \log H_\lambda'(y_{ij}),$$

where $\boldsymbol{z}_i = H_\lambda(y_i) - \boldsymbol{X}_i \boldsymbol{\beta}$. Since $\mathrm{E}[\boldsymbol{z}_i] = \mathbf{0}$, it follows that $\mathrm{E}[\partial^2 L / \partial \boldsymbol{\beta} \partial \tau^2] = \mathrm{E}[\partial^2 L / \partial \boldsymbol{\beta} \partial \sigma^2] = \mathbf{0}$. The other elements of the Fisher information can be obtained by a straightforward calculation. Moreover, under Assumptions 1 and 2, the each element of the Fisher information matrix is finite, so that the asymptotic normality of $\widehat{\boldsymbol{\phi}}$ follows.

A2. Proof of Theorem 2. Let $\boldsymbol{\phi}_0$ is the true values of parameters. It suffices to show that $P(\mu_i \le Q_a(y_i, \widehat{\boldsymbol{\phi}})) = a + O(m^{-1})$ for $a \in (0, 1)$. We first note that It holds that

$$P(\mu_i \le Q_a(y_i, \widehat{\boldsymbol{\phi}})) = \mathrm{E}[P(\mu_i \le Q_a(y_i, \widehat{\boldsymbol{\phi}}) | \boldsymbol{y}_s)] = \mathrm{E}[F(Q_a(y_i, \widehat{\boldsymbol{\phi}}); y_i, \boldsymbol{\phi}_0)],$$

where $F(\cdot; y_i, \boldsymbol{\phi}_0)$ is a distribution function of μ_i given y_i. Let $G(y_i, \widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}_0) = F(Q_a(y_i, \widehat{\boldsymbol{\phi}}); y_i, \boldsymbol{\phi}_0)$, noting that $0 \le G(y_i, \widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}_0) \le 1$ and $G(y_i, \boldsymbol{\phi}_0, \boldsymbol{\phi}_0) = a$. The Taylor expansion of $G(y_i, \widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}_0)$ shows that

$$\begin{aligned} G(y_i, \widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}_0) = G(y_i, \boldsymbol{\phi}_0, \boldsymbol{\phi}_0) &+ \sum_j G_{\phi_j}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} (\widehat{\phi}_j - \phi_j) \\ &+ \frac{1}{2} \sum_{j,k} G_{\phi_j \phi_k}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} (\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k) \\ &+ \frac{1}{6} \sum_{j,k,\ell} G_{\phi_j \phi_k \phi_\ell}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*} (\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)(\widehat{\phi}_\ell - \phi_\ell), \end{aligned}$$

where $\boldsymbol{\phi}^*$ is on the line connecting $\widehat{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}_0$. Then, it follows that

$$P(\mu_i \le Q_a(y_i, \widehat{\boldsymbol{\phi}})) = \mathrm{E}[G(y_i, \widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}_0)] = a + R_1 + \frac{1}{2} R_2 + \frac{1}{6} R_3,$$

where

$$R_1 = \mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)\right]$$
$$R_2 = \sum_{j,k} \mathrm{E}\left[G_{\phi_j\phi_k}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)\right]$$
$$R_3 = \sum_{j,k,\ell} \mathrm{E}\left[G_{\phi_j\phi_k\phi_\ell}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)(\widehat{\phi}_\ell - \phi_\ell)\right].$$

Using the Cauchy-Schwarz inequality, we have

$$\mathrm{E}\left[G_{\phi_j\phi_k}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)\right]$$
$$\leq \left\{\mathrm{E}[(\widehat{\phi}_j - \phi_j)^4]\right\}^{\frac{1}{4}} \left\{\mathrm{E}[(\widehat{\phi}_k - \phi_k)^4]\right\}^{\frac{1}{4}} \sqrt{\mathrm{E}\left[G_{\phi_j\phi_k}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)^2\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}\right]}.$$

From the asymptotic normality of $\widehat{\boldsymbol{\phi}}$ given in Theorem 1, it holds that $\mathrm{E}[|\widehat{\phi}_k - \phi_k|^r] = O(m^{-r/2})$. Moreover, since the range of $G(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)$ is $(0, 1)$, the partial derivatives of $G(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)$ are bounded. Then, we obtain $R_2 = O(m^{-1})$. Using the similar evaluation, we can show that $R_3 = O(m^{-1})$. Regarding R_1, it is noted that

$$\mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)\right] = \mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\mathrm{E}[\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0|y_i]\right].$$

From Lohr and Rao (2009), it holds $\mathrm{E}[\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0|y_i] = m^{-1}\boldsymbol{b}_{\boldsymbol{\phi}} - \boldsymbol{I}_{\boldsymbol{\phi}}^{-1}\partial L_i(y_i, \boldsymbol{\phi}_0)/\partial\boldsymbol{\phi} + o_p(m^{-1})$, where $\sum_{i=1}^m L_i(y_i, \boldsymbol{\phi}_0) \equiv L(\boldsymbol{\phi})$ and $b_{\boldsymbol{\phi}} = \lim_{m\to\infty} m\mathrm{E}[\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0]$ is the asymptotic bias of $\widehat{\boldsymbol{\phi}}$. Hence, we have

$$\mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\mathrm{E}[\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0|y_i]\right]$$
$$= \frac{1}{m}\mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\right]\boldsymbol{b}_{\boldsymbol{\phi}} - \mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\boldsymbol{I}_{\boldsymbol{\phi}}^{-1}\frac{\partial}{\partial\boldsymbol{\phi}}L_i(y_i; \boldsymbol{\phi}_0)\right] + o(m^{-1}),$$

which is $O(m^{-1})$. Therefore, the proof is completed.

A3. Proof of Theorem 3. From the proof of Theorem 2, we have

$$F_a(\boldsymbol{\phi}_0) \equiv P(\mu_i \leq Q_a(y_i, \widehat{\boldsymbol{\phi}})) = a + \frac{c_a(\boldsymbol{\phi}_0)}{m} + o(m^{-1}),$$

where $c_a(\boldsymbol{\phi})$ is a smooth function of $\boldsymbol{\phi}$. Let a^* and $\widehat{a}^*$ be satisfying $F_{a^*}(\boldsymbol{\phi}_0) = a$ and $F_{a^*}(\widehat{\boldsymbol{\phi}}) = a$, respectively. Then, from the above expansion, we have $\widehat{a}^* - a^* = o(m^{-1})$, thereby

$$P(\mu_i \leq Q_{\widehat{a}^*}(y_i, \widehat{\boldsymbol{\phi}})) = P(\mu_i \leq Q_{a^*}(y_i, \widehat{\boldsymbol{\phi}})) + o(m^{-1}) = a + o(m^{-1}),$$

which completes the proof.

A4. Checking assumptions of transformations. We here check the assumption 3 in Assumption 1 for the dual power (DP) transformation (10) and sinh-arcsinh (SS) transformation (11).

(DP transformation) We first note that $H_\lambda^{-1}(x) = O(x^{1/\lambda})$ as $x \to \infty$. By putting $x = -t$ for $t > 0$, we have

$$H_\lambda^{-1}(x) = (\sqrt{1 + \lambda^2 t^2} - \lambda t)^{1/\lambda} = \frac{1}{(\sqrt{1 + \lambda^2 t^2} + \lambda t)^{1/\lambda}} = O(t^{-1/\lambda})$$

as $t \to \infty$. A straightforward calculation shows that

$$\frac{\partial H_\lambda(x)}{\partial \lambda} = \frac{x^\lambda \log x + x^{-\lambda} \log x}{2\lambda} + \frac{x^\lambda - x^{-\lambda}}{2\lambda^2},$$

thereby, it follows that

$$\left| \frac{\partial H_\lambda}{\partial \lambda}(H_\lambda^{-1}(x)) \right| = O(|x| \log |x|) + O(|x|^{-1} \log |x|) + O(|x|) + O(|x|^{-1}) = O(|x| \log |x|)$$

as $|x| \to \infty$. Moreover, since

$$\frac{\partial^2 H_\lambda(x)}{\partial \lambda^2} = \frac{x^\lambda (\log x)^2 - x^{-\lambda} (\log x)^2}{2\lambda} - \frac{x^\lambda - x^{-\lambda}}{\lambda^3},$$

the similar evaluation leads to $\left| \partial^2 H_\lambda(w) / \partial \lambda^2 \right| = O(|x| (\log |x|)^2)$ as $|x| \to \infty$. Regarding $\partial^2 \log H'_\lambda(x) / \partial \lambda^2$, it holds that

$$\left| \frac{\partial^2 \log H'_\lambda(w)}{\partial \lambda^2} \right| = \left| \frac{4 (\log w)^2}{w^2 (w^{\lambda - 1} + w^{-\lambda - 1})^2} \right| = O((\log |x|)^2 |x|^2)$$

as $|x| \to \infty$, so that the DP transformation satisfies the assumption. When the location parameter is used, namely, $H_{\lambda,c}(x) = \{(x+c)^{\lambda} - (x+c)^{-\lambda}\}/2\lambda$, it is noted that $\partial^k H_{\lambda,c}(x)/\partial c^k = \partial^k H_{\lambda,c}(x)/\partial x^k$, so that the quite similar evaluation shows that the shifted-DP transformation also satisfies the assumption.

(SS transformation) It follows that

$$\frac{\partial H_{a,b}(x)}{\partial a} = -\cosh(b\sinh^{-1}(x) - a), \qquad \frac{\partial H_{a,b}(x)}{\partial b} = \cosh(b\sinh^{-1}(x) - a)\sinh^{-1}(x).$$

Note that $\sinh^{-1}(x) = O(\log|x|)$ as $|x| \to \infty$, so that $H_{a,b}^{-1}(x) = O(\exp(b^{-1}\log|x|)) = O(|x|^{1/b})$. Then, we have

$$\begin{aligned}
&\frac{\partial H_{a,b}}{\partial a}(H_{a,b}^{-1}(x)) = O(\exp(b\log|x|^{1/b})) = O(|x|),\\
&\frac{\partial H_{a,b}}{\partial b}(H_{a,b}^{-1}(x)) = O(\exp(b\log|x|^{1/b})\log|x|^{1/b}) = O(|x|\log|x|),
\end{aligned}$$

as $|x| \to \infty$. Moreover, it holds that

$$\begin{aligned}
&\frac{\partial^2 H_{a,b}(x)}{\partial^2 a} = \sinh(b\sinh^{-1}(x) - a), \qquad \frac{\partial^2 H_{a,b}(x)}{\partial^2 b} = \sinh(b\sinh^{-1}(x) - a)\{\sinh^{-1}(x)\}^2\\
&\frac{\partial^2 H_{a,b}(x)}{\partial a \partial b} = -\sinh(b\sinh^{-1}(x) - a)\sinh^{-1}(x),
\end{aligned}$$

thereby the similar evaluation shows that $\partial^2 H_{a,b}(x)/\partial^2 a = O(|x|)$, $\partial^2 H_{a,b}(x)/\partial^2 b = O(|x|(\log|x|)^2)$ and $\partial^2 H_{a,b}(x)/\partial a \partial b = O(|x|\log|x|)$ as $|x| \to \infty$. On the other hand, a straightforward calculation shows that

$$\frac{\partial}{\partial a}\log H'_{a,b}(x) = \frac{H_{a,b}(x)}{1 + H_{a,b}(x)^2}\frac{\partial H_{a,b}(x)}{\partial a}, \qquad \frac{\partial}{\partial b}\log H'_{a,b}(x) = \frac{1}{b} + \frac{H_{a,b}(x)}{1 + H_{a,b}(x)^2}\frac{\partial H_{a,b}(x)}{\partial b},$$

which are bounded by the function $\partial H_{a,b}(x)/\partial a$ and $\partial H_{a,b}(x)/\partial b$, respectively. A straightforward calculations show that the second partial derivatives of $\log H'_{a,b}(x)$ are bounded by polynomial functions of the second partial derivatives of $H_{a,b}(x)$ and $H_{a,b}(x)$, thereby the assumption is satisfied.

References

[1] Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28-36.

[2] Brent, R. (1973). *Algorithms for Minimization without Derivatives.* Englewood Cliffs N.J. Prentice-Hall.

[3] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformation (with discussion). *Journal of the Royal Statistical Society: Series B*, **26**, 211-252.

[4] Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014) Outliner robust small area estimation. *Journal of the Royal Statistical Society: Series B*, **76**, 47-69.

[5] Chatterjee, S., Lahiri, P. and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related predictions intervals in linear mixed models. *The Annals of Statistics*, **36**, 1221-1245.

[6] Fay, R. and Herriot, R. (1979). Estimators of income for small area places: an application of James–Stein procedures to census. *Journal of the American Statistical Association*, **74**, 341-353.

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

[8] Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B*, **68**, 221-238.

[9] Jiang, J. (1996). REML estimation: asymptotic behavior and related topics. *The Annals of Statistics*, **24**, 255-286.

[10] Jiang, J. (2010). *Large Sample Techniques for Statistics.* Springer.

[11] Jiang, J. and Lahiri, P. (2006). Estimation of finite population domain means: a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, **101**, 301-311.

[12] Jones, M. C. and Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, **96**, 761-780.

[13] Lahiri, P. and Mukherjee, K. (2007). On the design consistency property of hierarchical Bayes estimators in finite population sampling. *The Annals of Statistics*, **35**, 724-737.

[14] Li, Y, and Lahiri, P. (2007). Robust model-based and model-assisted predictors of the finite population total. *Journal of the American Statistical Association*, **102**, 664-673.

[15] Lohr, S. L. and Rao, J. N. K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika*, **96**, 457-468.

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

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

[18] Molina, I., Nandram, B. and Rao, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, **8**, 852-885.

[19] Nandram, B. (1999). An empirical Bayes prediction interval for the finite population mean of a small area. *Statistica Sinica*, **9**, 325-343.

[20] Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. and Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, B*, **70**, 265-286.

[21] Rao, J.N.K. and Molina, I. (2015) *Small Area Estimation, 2nd Edition*. Wiley.

[22] Schmid, T., Tzavidis, N., Münnich, R. and Chambers, R. (2016) Outlier robust small-area estimation under spatial correlation. *Scandinavian Journal of Statistics*, **43**, 806-826.

[23] Sugasawa, S. and Kubokawa, T. (2017). Transforming response values in small area prediction. *Computational Statistics & Data Analysis*, in press.

[24] Yang, Z. L. (2006). A modified family of power transformations. *Economics Letters*, **92**, 14-19.

[25] Yoshimori, M. and Lahiri, P. (2014). A second-order efficient empirical Bayes confidence interval. *The Annals of Statistics*, **42**, 1233-1261