Stat Papers
DOI 10.1007/s00362-017-0879-7

REGULAR ARTICLE

Small area estimation under transformed nested-error regression models

Huapeng Li[1,2] · **Yukun Liu**[1] · **Riquan Zhang**[1]

Received: 5 February 2016 / Revised: 10 January 2017

Abstract The empirical best linear unbiased prediction (EBLUP) based on the nested error regression model (Battese et al. in J Am Stat Assoc 83:28–36, 1988, NER) has been widely used for small area mean estimation. Its so-called optimality largely depends on the normality of the corresponding area level and unit level error terms. To allow departure from normality, we propose a transformed NER model with an invertible transformation, and employ the maximum likelihood method to estimate the underlying parameters of the transformed NER model. Motivated by Duan's (J Am Stat Assoc 78:605–610, 1983) smearing estimator, we propose two small area mean estimators depending on whether all the population covariates or only the population covariate means are available in addition to sample covariates. We conduct two design-based simulation studies to investigate their finite-sample performance. The simulation results indicate that compared with existing methods such as the empirical best linear unbiased prediction, the proposed estimators are nearly the same reliable when the NER model is valid and become more reliable in general when the NER model is violated. In particular, our method does benefit from incorporating auxiliary covariate information.

Keywords Empirical best linear unbiased prediction · (Adjusted) Empirical likelihood · Nested error regression model · Small area estimation · Transformed nested error regression model

✉ Yukun Liu
ykliu@sfs.ecnu.edu.cn

1 School of Statistics, East China Normal University, Shanghai 200241, People's Republic of China

2 School of Mathematics and Computer Sciences, Shanxi Datong University, Datong 037009, People's Republic of China

Published online: 06 February 2017

Mathematics Subject Classification Primary 62G05 · Secondary 62E20

1 Introduction

Motivated by the growing demand of reliable small area statistics in public and private sectors, the problem of small area estimation has received increasing attention and fruitful small area estimation techniques have been developed during the past decades. See, for example, the books by Rao (2003) and Rao and Molina (2015) for a comprehensive description or the paper by Pfeffermann (2013) for a thorough review. The term small area usually denotes any subpopulation for which direct estimates of adequate precision cannot be produced. Examples of small areas include geographical regions (such as states, counties) and socio-demographic groups (such as sex, age, race within a large geographic area). The primary aim of small area estimation is to make inferences with sufficient precision not only for the whole area under consideration but also for all sub-areas separately. Nevertheless, the sampling design of a typical national survey aims to only ensure that inferences can be made with sufficient precision for the whole area. This may result in few or even no sampling units in many sub-areas or small areas, which poses serious challenges to statisticians to estimate characteristics of the small areas with satisfactory precision.

The classical direct survey estimators for individual small areas are based only on area-specific sample data, and yield very large standard errors because of the shortage of data. It is widely accepted that this problem can be addressed by using indirect estimators, in particular, model-based small area estimators. Based on explicit small area models, model-based small area methods "borrow strength" from outside the small area, from the values of other variables in the small area, and from outside the survey, and produce reliable small area estimators.

Two celebrated models to achieve this purpose are the Fay–Herriot model (Fay and Herriot 1979) and the nested error regression model (Battese et al. 1988, NER). The well-accepted predictor/estimator of a small area mean is the Empirical Best Linear Unbiased Prediction (EBLUP). In the efforts of extending the NER model, Jiang and Nguyen (2012) proposed a heterogeneous NER model, which allowed heterogeneous errors across small areas. Their proposal was motivated by the varying variances across small areas of the Iowa Crops Data (See Fig. 1), which was first investigated by Battese et al. (1988). From Fig. 1, we also observe that the distributions of the response variable (sampled hectare), in particular for small areas 1, 5, 6 and 10, are severely skewed and clearly far from normal. This phenomenon is very common in practice since in practical survey sample, the response variable, such as income, revenue, harvest yield or production, is usually positive. In this situation, EBLUP will lose its optimality, which depends heavily on normality of the corresponding area level and unit level error terms.

Apart from robustifying (Schmid and Munnich 2014; Sinha and Rao 2009), a remedy to this issue is to transform the response variable so that the normality assumption is or is approximately met, and then model the transformed response variable by the NER model. Just as stated by Gurka et al. (2006), "transformation of the response has become a very simple and popular remedy in model fitting when the validity of

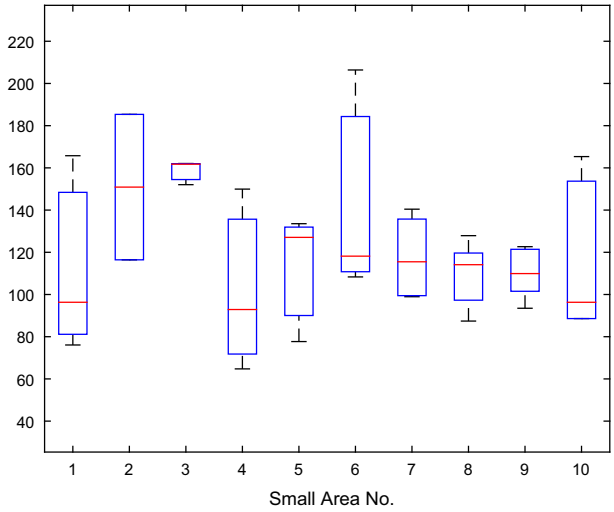

Fig. 1 Boxplots of the Iowa corn data

the assumptions of the model are called into question." The most commonly-used transformation is the Box–Cox transformation (Box and Cox 1964),

$$h_{\mathrm{BC}}(y;\lambda) = \begin{cases} (y^{\lambda}-1)/\lambda, & \text{if } \lambda \neq 0, \\ \log(y), & \text{if } \lambda = 0, \end{cases}$$

which requires $y > 0$ and was originally proposed in the linear regression model. Box and Cox (1964) also studied a shifted power transformation to avoid the positive response restriction to some extent. Gurka et al. (2006) extended it to the linear mixed model and explored its potential effect on estimation and inference of the model parameters. Unfortunately, since the range of the Box-Cox transformation is not the whole real line, the maximum likelihood estimator of the transformation parameter λ in is inconsistent (Sugasawa and Kubokawa 2014). Without the consistency of the estimated transformation, the estimation under the transformed model is questionable. Sugasawa and Kubokawa (2015) constructed a consistent estimator for this λ and Sugasawa and Kubokawa (2014) employed the dual power transformation (Yang 2006) and showed that the maximum likelihood estimator of the transformation parameter is consistent. However applications of their proposals are hindered by two facts: One is that they apply only to positive response variables; The other is that they are designed to make inference not for small area means themselves but for their transformations.

In this paper, we propose a new transformed NER model, where the transformation is simple and invertible, and its domain and range are both the whole real line. This offers much flexibility and makes it feasible for our subsequent small area estimation. The new transformation show some features similar to John and Draper (1980)'s modulus transformation, which was found to be appropriate for dealing with a fairly symmetric but non-normal error distribution. The small area estimation procedures

developed in this paper still work if our proposed transformation is replaced by the modulus transformation, or more generally, any transformation that is invertible on the whole real line and has certain smoothness in the underlying parameter. We decide for our version because of the concise forms of itself and its inverse. This not only facilitates its applications in practice but also makes it easy for us to prove the consistency of the maximum likelihood estimator (MLE) of the transformation parameter λ.

Besides λ, we propose to estimate the unknown parameters in the model by their MLEs, and establish their consistency. Motivated by Duan (1983)'s smearing estimator and based on the estimated transformation, we propose in Sect. 2 two small area mean estimators depending on whether the population covariates or only the population covariate means are available in addition to sample covariates. Our simulation results, given in Sect. 3, indicate that compared with EBLUP, the proposed estimator is as reliable when the NER model is valid and becomes more reliable when the NER model is violated; Furthermore, its performance can be remarkably improved by incorporating auxiliary covariate information. Section 4 provides a design-based simulation study based on a real finite population to further illustrate the usefulness and advantage of the proposed small area mean estimators. A discussion is given in Sect. 5. For clarity, all proofs are postponed to the appendix.

2 Transformed NER model

2.1 Model set-up

Suppose the population of interest consists of m subpopulations $\{(Y_{kj}, \mathbf{X}_{kj}) : \ j = 1, 2, \ldots, N_k\}_{k=1}^{m}$. Let the observed responses and the accompanying covariates be $\{(y_{kj}, \mathbf{x}_{kj}) : \ j = 1, 2, \ldots, n_k\}$, where n_k is the sample size in the kth area, and $k = 1, 2, \ldots, m$ are small area indices. In addition, it is often the case in practical survey that either all the covariate variables $\mathbf{X}_{kj}$'s or only the population covariate means $\bar{\mathbf{X}}_k$'s are available, which can be taken as auxiliary information. The primary goal of this paper is to estimate the population means $\bar{Y}_k$ of all the m small areas.

Battese et al. (1988) proposed to model the relationship between y_{kj} and $\mathbf{x}_{kj}$ by the NER model,

$$y_{kj} = \mathbf{x}_{kj}^{\top}\boldsymbol{\beta} + v_k + \varepsilon_{kj}, \tag{1}$$

where $\boldsymbol{\beta}$ is the regression coefficient, $v_k \overset{iid}{\sim} N(0, \sigma_v^2)$ denotes a random effect and $\varepsilon_{kj} \overset{iid}{\sim} N(0, \sigma_e^2)$ the random error. The random effects v_k's are assumed to be independent of ε_{kj}'s. The common regression coefficient $\boldsymbol{\beta}$ is used to borrow strength from other small areas, while the random effect v_k is used to explain the remaining variation of y that can not be explained by the covariate $\mathbf{x}$.

To allow possible departure of the NER model assumption, we may transform the response variable such that the NER model for the transformed response variable is or is approximately met (Gurka et al. 2006; Sugasawa and Kubokawa 2014, 2015).

Specifically, it is assumed that

$$h(y_{kj};\lambda) = \mathbf{x}_{kj}^{\top}\boldsymbol{\beta} + v_k + \varepsilon_{kj}, \tag{2}$$

where $h(y_{kj};\lambda)$ is a user-specified transformation and λ a tuning parameter to be determined by data. Gurka et al. (2006) suggested using the Box-Cox transformation, while Sugasawa and Kubokawa recommended using the dual power (DP) transformation (Yang 2006; Sugasawa and Kubokawa 2014),

$$h_{\mathrm{DP}}(y;\lambda) = \begin{cases} (y^{\lambda} - y^{-\lambda})/2\lambda, & \text{if } \lambda > 0, \\ \log(y), & \text{if } \lambda = 0, \end{cases}$$

and the dual power logistic (DPL) transformation (Sugasawa and Kubokawa 2015)

$$h_{\mathrm{DPL}}(y;\lambda) = \begin{cases} \{y^{\lambda}(1-y)^{-\lambda} - y^{-\lambda}(1-y)^{\lambda}\}/2\lambda, & \text{if } \lambda > 0, \\ \log\{y^{\lambda}(1-y)^{-\lambda}\}, & \text{if } \lambda = 0. \end{cases}$$

A common pitfall of the three transformations is that they restrict the response to be positive or belong to a real subset of the whole real line.

As a useful alternative transformation, we propose to use

$$h(y;\lambda) = \mathrm{Sign}(y) \times |y|^{\lambda}, \quad \lambda > 0, \tag{3}$$

where $\mathrm{Sign}(y)$ is defined to be -1, 0 and 1 if $y < 0$, $y = 0$ and $y > 0$, respectively. There are three considerations behind the choice of this transformation. First, since both v_k and ε_{kj} in (2) are normally distributed, the range of $h(y;\lambda)$ should be the whole real line. This excludes the Box-Cox, DP and DPL transformations, whose ranges are either $[-1/\lambda, \infty)$ or $(0, \infty)$ for any $\lambda > 0$. Both the domain and range of the new transformation are the whole real line. Second, $h(y;\lambda)$ should be invertible with respect to y, which is required in our proposed estimation procedure (See Sect. 2.3). The new transformation and John and Draper (1980)'s modulus transformation

$$h_{\mathrm{MOD}}(y;\lambda) = \begin{cases} \mathrm{Sign}(y)\{(|y|+1)^{\lambda} - 1\}/\lambda, & \text{if } \lambda \neq 0, \\ \mathrm{Sign}(y)\log(|y|+1), & \text{if } \lambda = 0, \end{cases}$$

both satisfy the two requirements. Our small area estimation procedures still work if we proceed with John and Draper (1980)'s modulus transformation in place of (3). More generally, our method extends to transformations that satisfy the above two requirements and are smooth enough with respect to the underlying parameter. Lastly, we choose transformation (3) to illustrate our main estimation procedure because its inverse is of a quite simple form

$$h^{-1}(y,\lambda) = \mathrm{Sign}(y)|y|^{1/\lambda}, \quad \lambda > 0, \tag{4}$$

which eases the proof burden of the consistency of the MLE of λ.

In transformation (3), we restrict the tuning parameter λ to be between $[c_1, c_2]$ for constants $c_1 \in (0, 1)$ and $c_2 \in (1, \infty)$. With this range of λ, transformation (3) includes the identity transformation as a special case. Hence the transformed NER model under (3) can be regarded as an extension of the NER model and provides more flexibility of model fitting. We would consequently expect better small area estimation under this model, compared with the EBLUP under the original NER model.

2.2 Parameter estimation

For ease of exposition, we denote $\mathbf{y}_k^{(\lambda)} = (y_{k1}^{(\lambda)}, \ldots, y_{kn_k}^{(\lambda)})^\top$ with $y_{kj}^{(\lambda)} = h(y_{kj}; \lambda)$, $\mathbf{x}_k = (\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kn_k})^\top$, and $\gamma = \sigma_v^2/\sigma_e^2$. The model assumptions imply that $\mathbf{y}_k^{(\lambda)} \sim N_{n_k}(\mathbf{x}_k\boldsymbol{\beta}, \Sigma_k)$, where $\Sigma_k = \sigma_e^2(\mathbf{I}_{n_k} + \gamma \mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top)$. The log-likelihood function (up to a constant not dependent on the unknown parameters) is

$$\ell_0(\lambda, \gamma, \sigma_e^2, \boldsymbol{\beta}) = -\frac{1}{2}\sum_{k=1}^{m}\{\log(|\Sigma_k|) + (\mathbf{y}_k^{(\lambda)} - \mathbf{x}_k\boldsymbol{\beta})^\top \Sigma_k^{-1}(\mathbf{y}_k^{(\lambda)} - \mathbf{x}_k\boldsymbol{\beta})\} + (\lambda - 1)\sum_{k=1}^{m}\sum_{j=1}^{n_k}\log(|y_{kj}|) + n\log(|\lambda|),$$

where $n = \sum_{k=1}^{m} n_k$ is the total sample size. It can be found that $\Sigma_k^{-1} = \mathbf{A}_k(\gamma)/\sigma_e^2$ where $\mathbf{A}_k(\gamma) = \mathbf{I}_{n_k} - \frac{\gamma}{1+\gamma n_k}\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top$, and $|\Sigma_k| = \sigma_e^{2n_k}(1 + n_k\gamma)$. Therefore the log-likelihood can be rewritten as

$$\ell_0(\lambda, \gamma, \sigma_e^2, \boldsymbol{\beta}) = -\frac{1}{2}\sum_{k=1}^{m}\left[\log(1 + n_k\gamma) + \frac{(\mathbf{y}_k^{(\lambda)} - \mathbf{x}_k\boldsymbol{\beta})^\top \mathbf{A}_k(\gamma)(\mathbf{y}_k^{(\lambda)} - \mathbf{x}_k\boldsymbol{\beta})\}}{\sigma_e^2}\right] - \frac{n}{2}\log(\sigma_e^2) + (\lambda - 1)\sum_{k=1}^{m}\sum_{j=1}^{n_k}\log(|y_{kj}|) + n\log(\lambda).$$

Given λ and γ, the MLEs of $\boldsymbol{\beta}$ and σ_e^2 are

$$\tilde{\boldsymbol{\beta}}(\lambda, \gamma) = \left\{\sum_{r=1}^{m}\mathbf{x}_r^\top \mathbf{A}_r(\gamma)\mathbf{x}_r\right\}^{-1}\sum_{s=1}^{m}\mathbf{x}_s^\top \mathbf{A}_s(\gamma)\mathbf{y}_s^{(\lambda)}, \tag{5}$$

$$\tilde{\sigma}_e^2(\lambda, \gamma) = \frac{1}{n}\sum_{k=1}^{m}\{\mathbf{y}_k^{(\lambda)} - \mathbf{x}_k\tilde{\boldsymbol{\beta}}(\lambda, \gamma)\}^\top \mathbf{A}_k(\gamma)\{\mathbf{y}_k^{(\lambda)} - \mathbf{x}_k\tilde{\boldsymbol{\beta}}(\lambda, \gamma)\}. \tag{6}$$

Putting them into $\ell_0(\lambda, \gamma, \sigma_e^2, \boldsymbol{\beta})$, we obtain the profile log-likelihood of (λ, γ),

$$\ell(\lambda, \gamma) = n \log(\lambda) + (\lambda - 1) \sum_{k=1}^{m} \sum_{j=1}^{n_k} \log(|y_{kj}|) - \frac{n}{2} \log \left\{ \tilde{\sigma}_e^2(\lambda, \gamma) \right\}$$
$$- \frac{1}{2} \sum_{k=1}^{m} \log(1 + n_k \gamma) - \frac{n}{2}. \tag{7}$$

Denote the maximum likelihood of (λ, γ) by $(\hat{\lambda}, \hat{\gamma}) = \arg\max_{\lambda, \gamma} \ell(\lambda, \gamma)$. Once $(\hat{\lambda}, \hat{\gamma})$ is obtained, we shall accordingly have the MLEs of $\boldsymbol{\beta}$ and σ_e^2,

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\lambda}, \hat{\gamma}) \quad \text{and} \quad \hat{\sigma}_e^2 = \tilde{\sigma}_e^2(\hat{\lambda}, \hat{\gamma}).$$

Remark 1 The main difficulty of the maximum likelihood estimation procedure is to calculate $(\hat{\lambda}, \hat{\gamma})$ or maximize $\ell(\lambda, \gamma)$. We solve this problem by a two-stage maximization method. Specifically, for fixed $\lambda \in (c_1, c_2)$, we calculate the profile log-likelihood $\ell_p(\lambda) = \max_{\gamma} \ell(\lambda, \gamma)$ and then obtain $\hat{\lambda}$ by maximizing $\ell_p(\lambda)$. Since the parameter λ is restricted to be positive in our proposed transformation given in (3), it is natural to choose c_1 and c_2 satisfying $0 < c_1 < c_2$. In the meantime, it would be desirable for transformation (3) to include the identity transformation as a special case. This suggests c_1 and c_2 lie in $(0, 1)$ and $(1, \infty)$, respectively. To make the family of the proposed transformations as large as possible, we recommend choosing c_1 to be as small as possible, and c_2 as large as possible. In our simulation study, we choose $c_1 = 0.001$ and $c_2 = 10$; Smaller c_1 or larger c_2 is also feasible.

The consistency of the MLEs of the unknown parameters are established in the following.

Theorem 1 *Assume the data* $(\mathbf{x}_{kj}, y_{kj})$*'s come from the transformed NER model* (2) *with the transformation given in* (3)*. Under conditions (C1)-(C3) in the appendix, the MLEs* $\hat{\lambda}$*,* $\hat{\boldsymbol{\beta}}$ *and* $\hat{\sigma}_e^2$ *are consistent estimators of* λ*,* $\boldsymbol{\beta}$ *and* σ_e^2*.*

The estimation of $\bar{Y}_k = \bar{X}_k \beta + v_k$ is equivalent to the estimation of a linear combination of β and the realization of the random effect v_k (Rao 2003, p. 80). Given the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\lambda}$, a prediction of the random effect v_k or an estimator of the realization of v_k under model (2) is

$$\hat{v}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} h(y_{kj}; \hat{\lambda}) - \bar{\mathbf{x}}_k^{\top} \hat{\boldsymbol{\beta}}.$$

In this paper we regard v_k as the realization of the random effect for fixed k and small area k $(1 \le k \le m)$. If the conditions of Theorem 1 are fulfilled, our proposed prediction/estimator $\hat{v}_k$ is conditionally consistent given v_k, the realization of the random effect, since $\hat{\lambda}$ and $\hat{\boldsymbol{\beta}}$ are both consistent.

In the next subsection, we shall present the proposed small area mean estimators, which are based on the estimated transformed NER model and the estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_e^2$

and $\hat{v}_k$. The consistencies of $\hat{\lambda}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_e^2$, and the conditional consistency of $\hat{v}_k$ make the proposed estimators valid.

2.3 Small area mean estimation

We begin by introducing Duan (1983)'s smearing estimator, which directly motivates our small area mean estimator. The smearing estimator was proposed for the mean of the untransformed response under a transformed linear regression model. Suppose $(y_i, \mathbf{x}_i)$ $(i = 1, 2, \ldots, n)$ is a simple sample from a transformed linear regression model,

$$g(y_i) = \mathbf{x}_i^\top \beta + e_i,$$

where $g(\cdot)$ is an invertible transformation and e_i's are independent and identically distributed with mean 0 and variance σ^2. The goal was to estimate the untransformed expectation $\mathbb{E}(y|\mathbf{x}_0) = \mathbb{E}\{g^{-1}(\mathbf{x}_0^\top \beta + e)\} = \int g^{-1}(\mathbf{x}_0^\top \beta + e) dF(e)$, where $F(\cdot)$ is the distribution of e. Given a consistent estimator $\hat{\beta}$ and the corresponding residuals $\hat{e}_i = g(y_i) - \mathbf{x}_i^\top \hat{\beta}$, Duan (1983) defined his smearing estimator for $\mathbb{E}(y_i|\mathbf{x}_0)$ as

$$\int g^{-1}(x_0\hat{\beta} + e) d\hat{F}_n(e) = \frac{1}{n}\sum_{i=1}^{n} g^{-1}(x_0\hat{\beta} + \hat{e}_i),$$

where $\hat{F}_n(e)$ is the empirical distribution of the residuals.

As discussed in Sect. 2.1, for $\lambda > 0$ the transformation $h(y; \lambda)$ in (3) is inverse with the inverse transformation given in (4). Model (2) can be equivalently written as

$$y_{kj} = h^{-1}(\mathbf{x}_{kj}^\top \boldsymbol{\beta}_0 + v_k + \varepsilon_{kj}; \lambda_0). \tag{8}$$

This implies that the small area mean in area k is

$$\begin{aligned} \bar{Y}_k &= \mathbb{E}\{h^{-1}(\mathbf{x}_{kj}^\top \boldsymbol{\beta}_0 + v_k + \varepsilon_{kj}; \lambda_0)|v_k\} \\ &= \int \left\{ \int h^{-1}(\mathbf{x}^\top \boldsymbol{\beta}_0 + v_k + t; \lambda_0) dF_\varepsilon(t) \right\} dF_{x,k}(\mathbf{x}) \\ &= \frac{1}{N_k}\sum_{j=1}^{N_k} \int h^{-1}(\mathbf{X}_{kj}^\top \boldsymbol{\beta}_0 + v_k + t; \lambda_0) dF_\varepsilon(t), \end{aligned} \tag{9}$$

where $F_\varepsilon(t)$ denotes the distribution function of ε and $F_{x,k}(\mathbf{x}) = N_k^{-1}\sum_{j=1}^{N_k} I(\mathbf{X}_{kj} \leq \mathbf{x})$ is the empirical distribution of $\{\mathbf{X}_{kj} : j = 1, 2, \ldots, N_k\}$. For vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, the inequality $\mathbf{x}_1 \leq \mathbf{x}_2$ holds element-wise.

Suppose for the time being that all the covariates $\mathbf{X}_{kj}$'s are known. Given $\hat{\sigma}_e$, a reasonable estimator for the distribution of ε is $N(0, \hat{\sigma}_e^2)$. Similar to Duan (1983)'s smearing estimator we proposed to estimate $\bar{Y}_k$ by

$$\hat{\bar{Y}}_k = \frac{1}{N_k}\sum_{j=1}^{N_k} \mathbb{E}_{\varepsilon^*}\{h^{-1}(\mathbf{X}_{kj}^\top \hat{\boldsymbol{\beta}} + \hat{v}_k + \varepsilon^*; \hat{\lambda})\},$$

where $\varepsilon^* \sim N(0, \hat{\sigma}_e^2)$ and $\mathbb{E}_{\varepsilon^*}$ denotes expectation with respect to the distribution of ε^*. If the expectation with respect to ε^* has no closed form, we propose to approximate it by a re-sampling method. Let $\varepsilon_1^*, \ldots, \varepsilon_B^*$ be independent observations generated from $N(0, \hat{\sigma}_e^2)$. Then by the weak law of large numbers

$$\mathbb{E}_{\varepsilon^*}\left\{h^{-1}(\mathbf{X}_{kj}^\top \hat{\boldsymbol{\beta}} + \hat{v}_k + \varepsilon^*; \hat{\lambda})\right\} \approx \frac{1}{B}\sum_{r=1}^{B} h^{-1}(\mathbf{X}_{kj}^\top \hat{\boldsymbol{\beta}} + \hat{v}_k + \varepsilon_r^*; \hat{\lambda}).$$

We require B to be a large integer, say 10,000, so that the above approximation is sufficiently precise. Finally the proposed small area estimator is

$$\hat{\bar{Y}}_k^{\text{TNER1}} = \frac{1}{BN_k}\sum_{j=1}^{N_k}\sum_{r=1}^{B} h^{-1}(\mathbf{X}_{kj}^\top \hat{\boldsymbol{\beta}} + \hat{v}_k + \varepsilon_r^*; \hat{\lambda}), \tag{10}$$

which is called TNER1 hereafter. The calculation of this estimator is very fast as it has a closed form.

Instead of knowing all covariates $\{\mathbf{X}_{kj}\}$, it is more often the case in practice that only the population means $\bar{\mathbf{X}}_k$ are available in addition to the sample covariates. To sufficiently incorporate the auxiliary information $\bar{\mathbf{X}}_k$, we propose to calibrate the estimator (10) by the empirical likelihood (Owen 1990, 2001) or adjusted empirical likelihood method (Chen et al. 2008). The resulting small area estimator, denoted TNER2, is

$$\hat{\bar{Y}}_k^{\text{TNER2}} = \sum_{j=1}^{n_k+1}\left\{\hat{p}_{kj}\cdot\frac{1}{B}\sum_{r=1}^{B} h^{-1}(\mathbf{x}_{kj}^\top \hat{\boldsymbol{\beta}} + \hat{v}_k + \varepsilon_r^*; \hat{\lambda})\right\}, \tag{11}$$

where $\hat{p}_{kj}$'s maximize the adjusted empirical likelihood function $\prod_{r=1}^{n_k+1} p_{kr}$ under the constraints

$$\sum_{r=1}^{n_k+1} p_{kr}(\mathbf{x}_{kr} - \bar{\mathbf{X}}_k) = 0, \quad p_{kr} \geq 0, \quad \sum_{r=1}^{n_k+1} p_{kr} = 1. \tag{12}$$

Here $\mathbf{x}_{k,n_k+1} = -\{1/(2n_k)\}(\bar{\mathbf{x}}_k - \bar{\mathbf{X}}_k) + \bar{\mathbf{X}}_k$ is an added pseudo-observation and used to guarantee the existence of feasible solutions to the constraints in (12). By the Lagrange multiplier method, it can be found that

$$\hat{p}_{kj} = \frac{1}{n_k+1}\frac{1}{1+\hat{\lambda}_k^\top(\mathbf{x}_{kj} - \bar{\mathbf{X}}_k)}, \quad j = 1, 2, \ldots, n_k+1, \tag{13}$$

where $\hat{\lambda}_k$ is the solution to $\sum_{j=1}^{n_k+1} \frac{\mathbf{x}_{kj} - \bar{\mathbf{X}}_k}{1+\lambda_k^\top(\mathbf{x}_{kj} - \bar{\mathbf{X}}_k)} = 0$.

The complicated form of the proposed estimators for $\bar{Y}_k$ makes it formidable to derive its limiting distribution. We make a remark on its conditional consistency instead. We can write the proposed TNER estimators in a unified form,

$$\hat{\bar{Y}}_k = \int \left\{ \int h^{-1}(\mathbf{x}^\top \hat{\boldsymbol{\beta}} + \hat{v}_k + t; \hat{\lambda}) d\hat{F}_{\varepsilon,*}(t) \right\} d\hat{F}_{x,k}(\mathbf{x}), \tag{14}$$

where $\hat{F}_{\varepsilon,*}(t)$ is the empirical distribution of $\varepsilon_1^*, \ldots, \varepsilon_B^*$ and $\hat{F}_{x,k}(\mathbf{x})$ is equal to $F_{x,k}(\mathbf{x})$ for TNER1 and equal to $\hat{F}_{x,k}^{(\mathrm{EL})}(\mathbf{x}) = \sum_{j=1}^{n_k+1} \hat{p}_{kj} I(\mathbf{x}_{kj} \le \mathbf{x})$ for TNER2. Roughly speaking, the conditional consistency of $\hat{\bar{Y}}_k$ will follow from the weak convergence of $\hat{F}_{\varepsilon,*}(t)$ and the consistency of $\hat{F}_{x,k}(\mathbf{x})$. Since $F_{x,k}(\mathbf{x})$ is a discrete distribution, the consistency of $\hat{F}_{x,k}(\mathbf{x})$ means that the finite number of estimated probability weights of $\hat{F}_{x,k}(\mathbf{x})$ converges in probability to those of $F_{x,k}(\mathbf{x})$. We have established in Theorem 1 the consistency of $\hat{\lambda}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_e^2$. As commented below Theorem 1, $\hat{v}_k$'s are conditionally consistent estimators of v_k's. By the Glivenko-Cantelli theorem, $\hat{F}_{\varepsilon,*}(t)$ converges uniformly to $N(0, \hat{\sigma}_e^2)$ conditionally on data as B becomes large. Since the consistency of $\hat{\sigma}_e^2$ implies the weak convergence of $N(0, \hat{\sigma}_e^2)$ to $N(0, \sigma_e^2)$, we conclude that $\hat{F}_{\varepsilon,*}(t)$ also converges weakly to $N(0, \sigma_e^2)$. Thus for each fixed $\mathbf{x}$ and conditionally on v_k, $\int h^{-1}(\mathbf{x}^\top \hat{\boldsymbol{\beta}} + \hat{v}_k + t; \hat{\lambda}) d\hat{F}_{\varepsilon,*}(t)$ is a conditionally consistent estimator of $\int h^{-1}(\mathbf{x}^\top \boldsymbol{\beta} + v_k + t; \hat{\lambda}) dF_{\varepsilon,*}(t)$ given $\mathbf{x}$ and v_k. The weak convergence of an empirical likelihood distribution estimator such as $\hat{F}_{x,k}^{(\mathrm{EL})}(\mathbf{x})$ has been established by Qin and Lawless (1994). Therefore both the small area mean estimators TNER1 and TNER2 are conditionally consistent.

3 Simulation study

3.1 Methods under comparison

This section provides simulation results to investigate the finite-sample performance of the proposed small area estimators under the transformed NER model. We have proposed two TNER estimators, depending on the accessibility of auxiliary information on covariate. It is often the case that population covariate means $\bar{\mathbf{X}}_k$'s are available, so we expect that TNER2 would be more promising. The study of TNER1 will provide us how much gain can be obtained if all covariates are available whereas we use only the population covariate means.

We compare TNER1 and TNER2 with three existing small area estimation methods. A naive estimator is the sample small area mean (direct method), which can be taken as a benchmark for comparison, although it is generally unreliable for small area estimation. As the second method, we may ignore model (2) and blindly use the EBLUP, which was proposed under the untransformed NER model (1). An EBLUP of the small area mean $\bar{Y}_k = \bar{\mathbf{X}}_k^\top \boldsymbol{\beta}_0 + v_k$ in the kth area under the NER model is

$$\hat{\bar{Y}}_k^{\mathrm{EBLUP}} = \bar{\mathbf{X}}_k^\top \tilde{\boldsymbol{\beta}} + \frac{n_k \tilde{\gamma}}{1 + n_k \tilde{\gamma}} (\bar{y}_k - \bar{\mathbf{x}}_k^\top \tilde{\boldsymbol{\beta}}), \tag{15}$$

where $\bar{y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} y_{kj}$, $\bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} \mathbf{x}_{kj}$ and $\tilde{\gamma}$ and $\tilde{\boldsymbol{\beta}}$ are the MLEs of γ and $\boldsymbol{\beta}$ under model (1). Here the covariate population means $\bar{\mathbf{X}}_k$ are assumed to be known. Given the MLEs $\hat{\boldsymbol{\beta}}$ and $\hat{v}_k$ of the transformed NER model (2), a natural estimator of the mean of the transformed response $\mu_k = \bar{\mathbf{X}}_k^\top \boldsymbol{\beta} + v_k$ is $\hat{\mu}_k = \bar{\mathbf{X}}_k^\top \hat{\boldsymbol{\beta}} + \hat{v}_k$. The third is Sugasawa and Kubokawa (2015)'s estimator (SS method), $h_{\mathrm{BC}}^{-1}(\hat{\mu}_k; \tilde{\lambda})$, with $\tilde{\lambda}$ an estimate of λ. This estimator is designed to estimate not the original small area means $\bar{Y}_k$ but $h_{\mathrm{BC}}^{-1}(\mu_k; \lambda)$. However $h_{\mathrm{BC}}^{-1}(\mu_k; \lambda)$ is generally different from $\bar{Y}_k$ unless the transformation is linear (Duan 1983). Although $h_{\mathrm{BC}}^{-1}(\hat{\mu}_k; \tilde{\lambda})$ may be a good estimation of $h_{\mathrm{BC}}^{-1}(\mu_k; \lambda)$, it generally has systematic bias when taken as an estimator of $\bar{Y}_k$.

3.2 Simulation settings

We begin by creating a finite population with $m = 16$ small areas, each consisting of 2,000 units. Then we draw random samples using simple random sampling without replacement from each area of the finite population; The sample sizes are chosen for convenience to be the same with $n_k = 5$ and 20. This procedure is repeated $R = 1000$ times and the Direct, TNER1, TNER2, EBLUP and SS estimates are calculated based on each sample. Let $\hat{\bar{Y}}_k^{(j)}$ denote a generic estimate of small area mean $\bar{Y}_k$ in the jth repetition. Two measures for the goodness of the estimator are considered: the relative root mean squared error (RRMSE) and the absolute relative bias (RBIAS) across all small areas,

$$\mathrm{RRMSE} = \frac{1}{m}\sum_{k=1}^{m} \mathrm{RRMSE}_k, \quad \mathrm{RBIAS} = \frac{1}{m}\sum_{k=1}^{m} \left| \frac{1}{\bar{Y}_k}\left(\frac{1}{R}\sum_{j=1}^{R} \hat{\bar{Y}}_k^{(j)} - \bar{Y}_k \right) \right|,$$

where $\mathrm{RRMSE}_k = \sqrt{\frac{1}{R}\sum_{j=1}^{R}(\hat{\bar{Y}}_k^{(j)} - \bar{Y}_k)^2}\big/\bar{Y}_k$. RRMSE reflects the overall performance of a small area estimation method. Clearly the smaller the RRMSE, the better the method. RBIAS is used to reflect whether a small area estimation method has systemic bias.

We generate the finite population from

$$h(y_{kj}; \lambda) = \mathbf{x}_{kj}^\top \beta + v_k + \varepsilon_{kj}, \quad k = 1, \ldots, 16. \tag{16}$$

The covariates $\mathbf{x}_{kj}$ are bivariate vectors with its first component 1 and each second component independently generated from the uniform distribution U(4, 8); The coefficient β is set to $(1, 2)^\top$. To show the effect of different values of λ, we consider three choices of λ_0: 0.3, 0.6 and 1. The random effects v_k and errors ε_{kj} are generated from three scenarios:

(I) both v_k and ε_{kj} are independently and identically distributed as $N(0, 1)$;
(II) for each k, v_k and ε_{kj} are independently generated from $N(0, 1)$ and $N(0, \sigma_k^2)$, respectively, and σ_k^2's are 0.2, 0.2, 0.2, 0.2, 0.4, 0.4, 0.4, 0.4, 0.6, 0.6, 0.6, 0.6, 2.0, 2.0, 2.0, 2.0;

(III) both v_k and ε_{kj} are independently generated from t-distribution with 3 degrees of freedom.

Here are some considerations behind the above simulation settings. Under scenario (I), when $\lambda_0 = 1$, the assumption of NER model is satisfied, and the EBLUP should have the best performance; We expect that the proposed TNER methods do not lose much efficiency. As λ_0 decreases from 1 to 0.6 and 0.3, the NER model is violated and the transformed NER model is satisfied; In this case we expect that the TNER methods outperform the EBLUP, SS and Direct methods. The rest scenarios are designed to study the robustness of the methods under comparison. In scenario (II), the homogeneity of the random errors are violated but they still follow normal distributions. The random effect and random error do no follow normal distributions any more in scenario (III), which is much heavier-tailed than normal.

3.3 Simulation results

The simulated RRMSE and RBIAS results of the small area estimation methods under consideration are tabulated in Table 1.

When $\lambda_0 = 1$ and in the case of scenario (I), the assumption of the EBLUP method is satisfied and EBLUP owns certain optimality. We find that the TNER1 and TNER2 estimators have almost the same RRMSE as EBLUP, while as expected the Direct estimator has the largest RRMSE. This indicates that the proposed TNER methods do not lose efficiency in the ideal setting of EBLUP, and it also necessitates the strategy of borrowing strength from other small areas.

When $\lambda_0 = 0.6$, all methods except the Direct method have very close RRMSEs. As λ_0 decreases to 0.6 and 0.3, the priority of TNER1 and TNER2 over SS, Direct and EBLUP becomes more and more obvious. For example, when $\lambda_0 = 0.3$, TNER1 has remarkable RRMSE reductions (more than 25%) compared with EBLUP and SS. To our surprise, TNER2 has very close performance compared with TNER1 although it uses not the information of all population covariates but only that of sample covariates and population covariate means.

When the TNER model is violated in the rest scenarios, the priority order of all methods keeps unchanged, namely the TNER methods are still the winners. All methods have reduced RRMSEs in scenario (II) and inflated RRMSEs in scenario (III). A possible reason is that compared with the unit error variance in scenario (I), the error variances of most small areas in scenario (II) are much smaller than 1, while those in scenario (III) are 3, much larger.

We turn to examining the RBIAS results. In all cases, it can be seen that the Direct method always has the smallest and negligible bias because it is unbiased in theory. When $\lambda_0 = 1$, in the case scenario (I), the assumption of EBLUP is satisfied, the rest four methods still have almost the same performance. However, as λ_0 decreases to $\lambda_0 = 0.6$ and 0.3, compared with EBLUP, the RBIAS of the TNER methods becomes less, while that of the SS method becomes much larger. there are similar results in the case scenario (II) and (III). As the sample size n_k per small area increases from 5 to 20, just like all RRMSEs decrease, we find that RBIAS of the TNER1, TNER2 and EBLUP also decrease, however, the SS method has increased RBIAS.

Table 1 RRMSE (unit: 0.1) and RBIAS (unit: 0.01) results of the five small area mean estimators under consideration

n_k	$\lambda = 0.3$					$\lambda = 0.6$					$\lambda = 1$				
	SS	Direct	TNER1	TNER2	EBLUP	SS	Direct	TNER1	TNER2	EBLUP	SS	Direct	TNER1	TNER2	EBLUP
Scenario I															
RRMSE															
5	1.49	2.26	1.04	1.12	1.38	0.57	1.37	0.54	0.54	0.56	0.33	0.84	0.33	0.33	0.31
20	1.33	1.32	0.52	0.57	0.71	0.33	0.68	0.26	0.28	0.28	0.17	0.42	0.17	0.17	0.17
RBIAS															
5	11.53	0.81	2.13	2.88	3.61	1.73	0.41	1.10	1.07	1.20	0.69	0.25	0.66	0.67	0.69
20	12.18	0.31	0.54	0.64	0.98	1.97	0.17	0.28	0.28	0.31	0.18	0.11	0.18	0.18	0.18
Scenario II															
RRMSE															
5	1.39	2.57	0.89	1.00	1.21	0.49	1.35	0.46	0.46	0.47	0.28	0.82	0.28	0.28	0.28
20	1.26	1.29	0.45	0.50	0.62	0.30	0.67	0.23	0.23	0.24	0.14	0.41	0.14	0.14	0.14
RBIAS															
5	11.23	0.81	2.08	2.96	3.25	1.62	0.40	0.94	0.90	1.03	0.60	0.24	0.55	0.57	0.59
20	11.82	0.32	1.32	1.20	0.87	1.87	0.18	0.34	0.33	0.26	0.16	0.12	0.15	0.15	0.15
Scenario III															
RRMSE															
5	2.08	3.48	1.80	1.81	2.65	0.91	1.61	0.91	0.90	0.96	0.56	1.00	0.58	0.57	0.57
20	1.71	1.71	0.88	0.90	1.41	0.50	0.80	0.45	0.45	0.48	0.28	0.49	0.29	0.29	0.29
RBIAS															
5	13.96	0.85	3.49	3.66	7.65	2.23	0.41	1.63	1.57	2.13	1.19	0.25	0.84	0.92	1.22
20	15.19	0.52	1.23	1.32	2.33	2.22	0.26	0.42	0.40	0.51	0.42	0.16	0.15	0.16	0.28

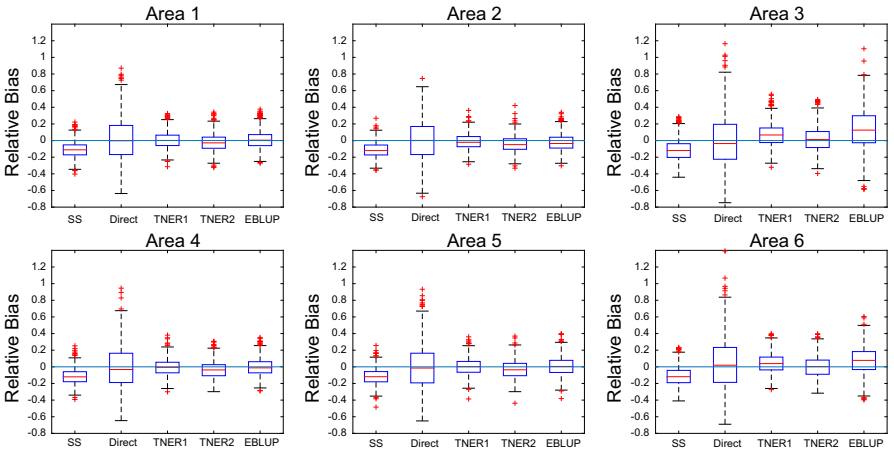

Fig. 2 Boxplots of the relative biases of the SS, Direct, TNER1, TNER2 and EBLUP mean estimates for areas 1–6 in scenario (I) with $\lambda = 0.3$ and $n_k = 5$

To see more clear the advantages of the TNER1 and TNER2 methods over the rest three methods, we plot the boxplots of the relative biases of the five estimators under comparison for each area in scenario (I) with $\lambda_0 = 0.3$ and $n_k = 5$. Figure 2 displays the boxplots for the first six areas. Those for the last ten areas are similar and omitted. Our general observations are the same as above. The SS predictor is clearly biased and the direct estimator has the largest variance although it is unbiased. The relative biases of the TNER1, TNER2 and EBLUP predictors are generally very small. In almost all areas, both the TNER1 and TNER2 predictors have smaller variances than EBLUP. In particular, in areas 3 and 6, EBLUP has not only larger biases but also larger variances than TNER1 and TNER2.

4 Empirical studies

4.1 A real data-set

To be more practical than the simulation settings in the previous section, we conduct simulations based on the data in the Survey of Labour and Income Dynamics (SLID) provided by Statistics Canada (2014). We take the survey data as a basis to create a realistic finite population and examine how well our proposed small area mean estimator and its competitors perform if we sample from this "real" population. Since the requirement of TNER1 is too strong, we do not take it into account and study TNER2 instead.

The data we obtained contains 147 variables and 47705 sampling units. We keep 6 of the 147 variables, i.e., `ttin`, `gender`, `age`, `yrx`, `tweek` and `edu`, standing respectively for: total income, gender, age, years of experience, number of weeks employed and education level. More precise definitions are not essential here. We keep the variable `ttin` because at the heart of the survey's objectives is the understanding of the

economic well-being of Canadians, gender and age are natural social-demographic variables that are used to construct areas for small area estimation studies. The rest three variables yrx, tweek and edu are believed to be closely related to ttin, and hence are taken as covariates. We remove any units containing missing values in these 6 variables as well as those with ttin ≤ 0 or the ages under 25. Negative ttin values also stand for missing data and most people with age < 25 are part-time or part-year workers. The resulting data set still contains 30,099 sampling units. We first create 8 age groups formed by individuals whose ages are in following intervals:

[25,30)	[30,35)	[35,40)	[40,45)	[45,50)	[50,55)	[55,60)	[60,∞)

Each age group is then divided into male and female sub-populations. Subsequently, we created a finite population with 16 small areas based on age-gender combinations. The sizes of these small areas N_k are given as follows.

Male	1372	1337	1469	1536	1866	1890	1920	3089
Female	1449	1504	1497	1695	2053	2019	1944	3459

4.2 Model and diagnostics

We take total income ttin as a response variable and yrx, tweek and edu as covariates. Suppose we have a random sample of size $n = 16n_k$ from the population. The goal is to estimate the average total incomes for all the 16 sub-populations. In addition to the sample covariates, we assume that all population covariate means are known. Hence not TNER1 but TNER2 applies. We fit the data by a transformed NER model using transformation (3), namely,

$$h(\texttt{ttin}_{kj};\lambda) = \beta_0 + \beta_1\,\texttt{yrx}_{kj} + \beta_2\,\texttt{tweek}_{kj} + \beta_3\,\texttt{edu}_{kj} + v_k + \varepsilon_{kj}. \quad (17)$$

For comparison, we also model the data by the usual NER model,

$$\texttt{ttin}_{kj} = \beta_0 + \beta_1\,\texttt{yrx}_{kj} + \beta_2\,\texttt{tweek}_{kj} + \beta_3\,\texttt{edu}_{kj} + v_k + \varepsilon_{kj}. \quad (18)$$

If the (transformed) NER model is valid, then the residuals are approximately equal to the corresponding random errors, which follow a normal distribution. Hence we can check the validation by studying whether the distribution of the residuals is close to normal. Figure 3 displays QQ-plots of the residuals under models (17) and (18) based on one random sample. It is clear that the QQ-plot based on model (17) is nearly a straight line, while that based on model (18) is severely not. Hence the residuals based on model (17) approximately follow a normal distribution but those based on model (18) is far from normal. This indicates that the transformed NER model is more appropriate than the original NER model.

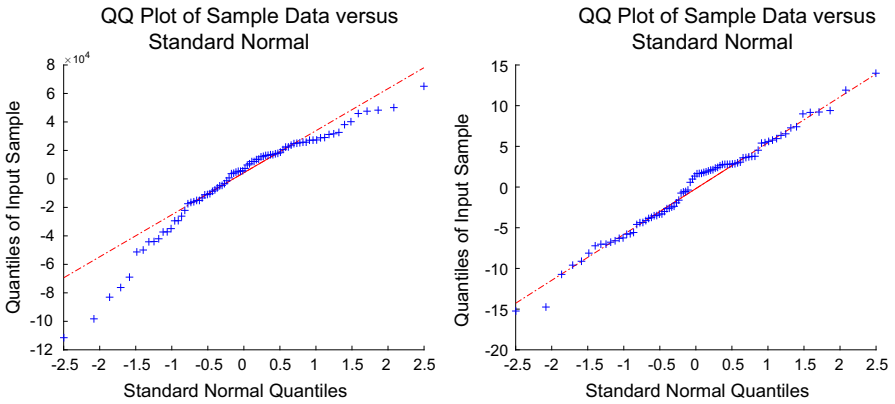

Fig. 3 QQ-plots of the residuals of the transformed NER model in (17) and the NER model in (18) when $n_k = 5$

We also conduct formal goodness-of-fit tests for the normality of the residuals by the Shapiro-Wilk test. For model (18), the p values is 4.12×10^{-4}, which is far less than 0.01 and strongly rejects the normality of the residuals and the appropriateness of the NER model. In comparison, the p values for model (17) is 0.52, much larger than 0.01. Hence there is no evidence for the non-normality of the residuals based on model (17). These results strongly confirm that the transformed NER model is appropriate for the sample while the NER model is not. This phenomenon occurs for around 62% of our simulated 1000 samples in the next subsection. Around 7% of the samples support both models. Even so, there are still 31% of the samples, in which neither model is appropriate.

4.3 Simulation results

We proceed to conduct simulations and consider three sample sizes $n_k = 5$, 10 and 20. Other than the finite population and the sample size, this simulation has every other aspects unchanged from the last section.

The simulation results are tabulated in Table 2. In this case, neither the untransformed NER nor the transformed NER model is correct. We find that the comparison results of the TNER2 with the SS, Direct, and EBLUP methods are very close to those in the simulation section. The TNER2 method outperforms the rest methods as it generally has the smallest RRMSE. When $n_k = 5$ and 10, the EBLUP follows, the SS method ranks the third, and the direct method is the most unstable. When $n_k = 20$, the EBLUP not only catches up with the TNER2, but also is less biased, while the SS method is outperformed by the Direct method. As the sample size n_k increases, all numbers of RRMSE and RBIAS decrease except for the RBIAS of the SS method. This is because the SS method suffers from systematic bias, as disclosed in our simulation study.

To get more insights into the results, we display the plots of RRMSE_k versus small area number k in Fig. 4. A lower line indicates smaller RRMSE_k and better overall

Table 2 RRMSE (unit: 0.1) and RBIAS (unit: 0.01) results of the real-data based simulation when all covariate means $\bar{\mathbf{X}}_k$'s are available in addition to the sample covariates

	n_k	SS	Direct	TNER2	EBLUP
RRMSE	5	2.12	3.63	1.73	1.96
	10	1.97	2.53	1.55	1.62
	20	1.89	1.79	1.39	1.31
RBIAS	5	16.65	0.84	10.47	11.94
	10	17.03	0.64	10.39	10.78
	20	17.07	0.41	9.21	8.52

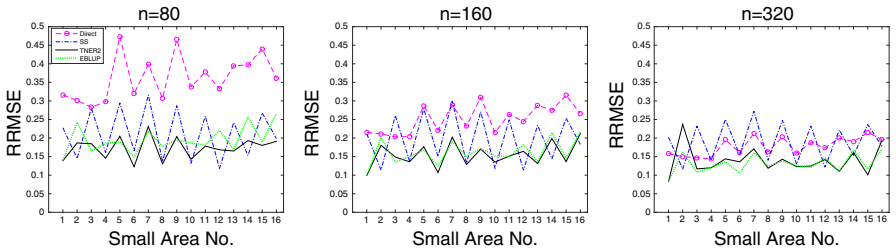

Fig. 4 RRMSE of different small area mean estimators when $n = 80$, 160 and 320

performance. When the sample size is small such as $n_k = 5$ or 10, it can be seen that the Direct estimator usually has the largest mean square errors and is the most unstable in most of small areas. By contrast, the SS estimator is better ($n_k = 5$) or comparable ($n_k = 10$) and TNER2 and EBLUP are much better. It is worth pointing out that TNER2 has even smaller RRMSEs than EBLUP for almost all small areas. When the sample size n_k increases to 20, the Direct estimator outperforms SS across almost all small areas, indicating that SS is unacceptable any more. In this situation, TNER2 and EBLUP have almost equally small RRMSEs.

Overall, the proposed small area mean estimation method produces reliable estimates in general and exhibits certain priority over the Direct, SS and EBLUP methods, in particular for small sample sizes. When the sample size is large such as $n_k = 20$, EBLUP is still a good alternative.

5 Discussion

To alleviate the dependence of the EBLUP on the NER model, we propose a new transformed NER model with an invertible transformation to model small area data. We establish the consistency of the MLEs of the transformation index parameters. Borrowing the idea of Duan (1983)'s smearing estimator, we propose two small area mean estimators under the new transformed NER model, depending on the extent to which auxiliary covarate information is available. Our simulation results provide evidence for the priority of the proposed small area mean estimator over existing estimators such as EBLUP when sample sizes are small.

The new estimation methodology also applies to more general transformations than (3) if they are invertible and smooth enough. In practice, we may fit data by

the untransformed NER model and check its goodness-of-fit by QQ-plots or a formal normality test of the residuals. We shall use the EBLUP method to estimate small area means if the normality of the residuals is not rejected. Otherwise, we proceed with the transformed NER model and apply the proposed small area mean estimators. We would recommend the use of transformation (3) although John and Draper (1980)'s modulus transformation or other transformations are also applicable.

In this paper our focus is on point estimation of small area means. It would be preferable to construct confidence intervals and evaluate the goodness of small area mean estimators, which necessitates the estimation of mean square errors of each small area mean estimator. There have been extensive studies in the literature on the estimation of prediction mean squared error under a basic unit-level model. See, for example, Prasad and Rao (1990), Datta and Lahiri (2000), Pfeffermann and Correa (2012), and Rao and Molina (2015). Due to the complexity and difficulty of small area estimation, resampling methods are often employed to estimate prediction mean squared error, such as a double-bootstrap procedure (Hall and Maiti 2006) and a parametric bootstrap method (González-Manteiga et al. 2008). This issue becomes more challenging under a transformed NER model because the transformation index parameter λ needs also to be estimated, which makes the property of the resulting small area estimator formidable. A possible solution is to employ a parametric bootstrap method as suggested by an anonymous referee. We would leave this as a further research topic.

Acknowledgements We thank the editor, an associate editor, and two anonymous referees for their constructive suggestions that significantly improved the paper. The research was supported in part by National Natural Science Foundation of China (Grant Numbers 11371142, 11501354, and 11571112), Program of Shanghai Subject Chief Scientist(14XD1401600), the 111 Project (B14019), and Doctoral Fund of Ministry of Education of China (20130076110004).

Appendix

To study the consistency of the MLEs, we make the following assumptions.

(C1) There exists constants $0 < c_1 < c_2$ such that the transformed response $h(y; \lambda_0)$ with $\lambda_0 \in [c_1, c_2]$ and the transformation in (3) satisfy the NER model in (1).
(C2) The small area number m keeps fixed and that as n goes to infinity, $n_k/n = \rho_k + o(1)$ for $\rho_k \in (0, 1)$.
(C3) Assume that $\Sigma_x = \sum_{k=1}^{m} \rho_k \Sigma_{xk}$ is nonsingular, where Σ_{xk} is the variance matrix of $\mathbf{x}_{kj}$ in small area k.

Apparently condition (C1) requires y_{kj} is an increasing function of $\mathbf{x}_{kj}^\top \beta + v_k + \varepsilon_{kj}$ since $h(y; \lambda_0)$ is increasing when $\lambda_0 > 0$. If y_{kj} is an decreasing function of $\mathbf{x}_{kj}^\top \beta + v_k + \varepsilon_{kj}$, it must be an increasing function of $\mathbf{x}_{kj}^\top(-\beta) + (-v_k) + (-\varepsilon_{kj})$. Then condition (C1) is still satisfied except that the parameter β has an opposite sign. Condition (C2) is imposed to provide a justification of the proposed estimation method for λ. Also we can check whether there exists system bias in the proposed method although the sample sizes n_k are generally very small in the literature of small area estimation.

Proof of Theorem 1 We begin by proving the consistency of $\hat{\lambda}$. Let (λ_0, γ_0) be the true value of (λ, γ). It is sufficient to show that as n is large, (1)

$$S = \frac{1}{n}\frac{\partial \ell(\lambda_0, \gamma_0)}{\partial \lambda} = o_p(1) \tag{19}$$

and (2) $\frac{1}{n}\frac{\partial^2 \ell(\lambda_0,\gamma_0)}{\partial \lambda \partial \lambda^\top}$ is positive definite. We shall prove only (1) since (2) can be proved along the same line of proving (1) but with more tedious derivation.

From (7), we have

$$S = \frac{1}{n}\frac{\partial \ell(\lambda_0, \gamma_0)}{\partial \lambda} = \frac{1}{\lambda_0} + \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}\log(|y_{kj}|) - \frac{1}{2}\frac{\partial \tilde{\sigma}_e^2(\lambda_0, \gamma_0)/\partial\lambda}{\tilde{\sigma}_e^2(\lambda_0, \gamma_0)}.$$

To simplify S, we need to investigate $\tilde{\sigma}_e^2(\lambda_0, \gamma_0)$ and $\partial\tilde{\sigma}_e^2(\lambda_0, \gamma_0)/\partial\lambda$. It follows from $\mathbf{A}_k(\gamma) = \mathbf{I}_{n_k} - \frac{\gamma}{1+\gamma n_k}\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top$ that for any fixed $\gamma > 0$,

$$\begin{aligned}
\frac{1}{n_r}\mathbf{x}_r^\top \mathbf{A}_r(\gamma)\mathbf{x}_r &= \frac{1}{n_r}\sum_{j=1}^{n_r}\mathbf{x}_{rj}\mathbf{x}_{rj}^\top - \frac{n_r\gamma}{1+\gamma n_r}\bar{\mathbf{x}}_r\bar{\mathbf{x}}_r^\top \\
&= \frac{1}{n_r}\sum_{i=1}^{n_r}(\mathbf{x}_{rj} - \bar{x}_r)(\mathbf{x}_{rj} - \bar{x}_r)^\top + \frac{\bar{\mathbf{x}}_r\bar{\mathbf{x}}_r^\top}{1+n_r\gamma} \\
&= \Sigma_{xr} + O_p(n^{-1/2}),
\end{aligned}$$

$$\begin{aligned}
\frac{1}{n_r}\mathbf{x}_r^\top \mathbf{A}_r(\gamma)\mathbf{y}_r^{(\lambda_0)} &= \frac{1}{n_r}\sum_{j=1}^{n_r}\mathbf{x}_{rj}y_{rj}^{(\lambda_0)} - \frac{n_r\gamma}{1+\gamma n_r}\bar{\mathbf{x}}_r\bar{y}_r^{(\lambda_0)} \\
&= \frac{1}{n_r}\sum_{j=1}^{n_r}(\mathbf{x}_{rj} - \bar{x}_r)(y_{rj}^{(\lambda_0)} - \bar{y}_r^{(\lambda_0)}) + \frac{\bar{\mathbf{x}}_r\bar{y}_r^{(\lambda_0)}}{1+\gamma n_r} \\
&= \Sigma_{xr}\beta_0 + O_p(n^{-1/2}),
\end{aligned}$$

$$\begin{aligned}
\frac{1}{n_r}\{\mathbf{y}_r^{(\lambda)}\}^\top \mathbf{A}_r(\gamma)\mathbf{y}_r^{(\lambda)} &= \frac{1}{n_r}\sum_{j=1}^{n_r}\{y_{rj}^{(\lambda)}\}^2 - \frac{n_r\gamma}{1+\gamma n_r}\{\bar{y}_r^{(\lambda)}\}^2 \\
&= \frac{1}{n_r}\sum_{j=1}^{n_r}\{(\mathbf{x}_{rj} - \bar{\mathbf{x}}_r)^\top\beta_0 + (\varepsilon_{rj} - \bar{\varepsilon}_r)\}^2 + O_p(n^{-1}) \\
&= \beta_0^\top\Sigma_{xr}\beta_0 + \sigma_e^2 + O_p(n^{-1/2}).
\end{aligned}$$

The above three equalities imply that

$$\frac{1}{n}\sum_{r=1}^{m}\mathbf{x}_r^\top \mathbf{A}_r(\gamma)\mathbf{x}_r = \sum_{r=1}^{m}\rho_r\Sigma_{xr} + O_p(n^{-1/2}), \tag{20}$$

$$\frac{1}{n}\sum_{r=1}^{m}\mathbf{x}_r^\top\mathbf{A}_r(\gamma)\mathbf{y}_r^{(\lambda_0)}=\sum_{r=1}^{m}\rho_r\Sigma_{xr}\beta_0+O_p(n^{-1/2}),\tag{21}$$

$$\frac{1}{n}\sum_{r=1}^{m}\{\mathbf{y}_r^{(\lambda)}\}^\top\mathbf{A}_r(\gamma)\mathbf{y}_r^{(\lambda)}=\beta_0^\top\sum_{r=1}^{m}\rho_r\Sigma_{xr}\beta_0+\sigma_e^2+O_p(n^{-1/2}).$$

By these three equalities, we immediately have

$$\tilde{\sigma}_e^2(\lambda_0,\gamma_0)=\sigma_e^2+O_p(n^{-1/2}).\tag{22}$$

To calculate $\partial\tilde{\sigma}_e^2(\lambda_0,\gamma_0)/\partial\lambda$, we denote $z_{kj}\equiv\partial y_{rj}^{(\lambda_0)}/\partial\lambda=y_{rj}^{(\lambda_0)}\log(|y_{kj}|)$. Then

$$\begin{aligned}\frac{\partial}{\partial\lambda}\frac{1}{n}\sum_{s=1}^{m}\mathbf{x}_s^\top\mathbf{A}_s(\gamma)\mathbf{y}_s^{(\lambda_0)}&=\sum_{r=1}^{m}\frac{n_r}{n}\Big[\frac{1}{n_r}\sum_{j=1}^{n_r}(\mathbf{x}_{rj}-\bar{x}_r)(y_{rj}^{(\lambda)}-\bar{y}_r^{(\lambda)})+\frac{\bar{\mathbf{x}}_r\bar{y}_r^{(\lambda)}}{1+\gamma n_r}\Big]\\&=\sum_{r=1}^{m}\rho_r\mathbb{C}\mathrm{ov}(\mathbf{x}_{rj},z_{rj})+O_p(n^{-1/2})\\&=\Sigma_{xz}+O_p(n^{-1/2}),\\\frac{\partial}{\partial\lambda}\frac{1}{n}\sum_{r=1}^{m}\{\mathbf{y}_r^{(\lambda)}\}^\top\mathbf{A}_r(\gamma)\mathbf{y}_r^{(\lambda)}&=\sum_{r=1}^{m}\rho_r\{2\beta_0^\top\mathbb{C}\mathrm{ov}(\mathbf{x}_{rj},z_{rj})+2\mathbb{C}\mathrm{ov}(z_{rj},\varepsilon_{rj})\}\\&=2\beta_0^\top\Sigma_{xz}+2\Sigma_{z\varepsilon}+O_p(n^{-1/2}).\end{aligned}$$

It can be found that

$$\begin{aligned}\tilde{\sigma}_e^2(\lambda,\gamma)=&\frac{1}{n}\sum_{k=1}^{m}\{\mathbf{y}_k^{(\lambda)}\}^\top\mathbf{A}_k(\gamma)\mathbf{y}_k^{(\lambda)}\\&-\{\frac{1}{n}\sum_{s=1}^{m}\mathbf{x}_s^\top\mathbf{A}_s(\gamma)\mathbf{y}_s^{(\lambda)}\}^\top\left\{\frac{1}{n}\sum_{r=1}^{m}\mathbf{x}_r^\top\mathbf{A}_r(\gamma)\mathbf{x}_r\right\}^{-1}\frac{1}{n}\sum_{s=1}^{m}\mathbf{x}_s^\top\mathbf{A}_s(\gamma)\mathbf{y}_s^{(\lambda)}.\end{aligned}$$

Accordingly

$$\begin{aligned}\frac{\partial}{\partial\lambda}\tilde{\sigma}_e^2(\lambda_0,\gamma_0)=&2\frac{1}{n}\sum_{k=1}^{m}\{\frac{\partial}{\partial\lambda}\mathbf{y}_k^{(\lambda_0)}\}^\top\mathbf{A}_k(\gamma_0)\mathbf{y}_k^{(\lambda_0)}-2\{\frac{1}{n}\sum_{s=1}^{m}\mathbf{x}_s^\top\mathbf{A}_s(\gamma_0)\frac{\partial}{\partial\lambda}\mathbf{y}_s^{(\lambda_0)}\}^\top\\&\times\left\{\frac{1}{n}\sum_{r=1}^{m}\mathbf{x}_r^\top\mathbf{A}_r(\gamma_0)\mathbf{x}_r\right\}^{-1}\frac{1}{n}\sum_{s=1}^{m}\mathbf{x}_s^\top\mathbf{A}_s(\gamma_0)\mathbf{y}_s^{(\lambda_0)}+o_p(1)\\=&2\beta_0^\top\Sigma_{xz}+2\Sigma_{z\varepsilon}-2\beta_0^\top\Sigma_x\Sigma_x^{-1}\Sigma_{xz}+o_p(1)\\=&2\Sigma_{z\varepsilon}+o_p(1).\end{aligned}\tag{23}$$

Putting (23) into (19), we obtain

$$S = \frac{1}{\lambda_0} + \sum_{k=1}^{m} \rho_k \mathbb{E}\{\log(|y_{kj}|)\} - \frac{1}{2}\frac{\Sigma_{z\varepsilon}}{\sigma_e^2} + o_p(1).$$

To prove $S = o_p(1)$, it is sufficient to prove for $r = 1, 2, \ldots, m$ that

$$\frac{1}{\lambda_0} + \mathbb{E}\{\log(|y_{rj}|)\} - \frac{1}{\sigma_e^2}\mathbb{C}\text{ov}(z_{rj}, \varepsilon_{rj}) = 0,$$

which is true as shown in Lemma 1. This proves the consistency of $\hat{\lambda}$.

Note that Eqs. (20) and (21) holds for any $\gamma > 0$. Since $\hat{\lambda}$ is consistent, we immediately obtain that $\hat{\beta}$ is an consistent estimator of β_0. By re-studying the proof of (22), we find that it is still true when γ_0 is replaced by any positive γ and λ_0 is replaced by its consistent estimator $\hat{\lambda}$. This completes the proof of Theorem 1. □

Lemma 1 *Under the assumptions for the transformed NER model, it holds that*

$$\frac{1}{\lambda_0} + \mathbb{E}_{x,\varepsilon}\left\{\log(|y_{kj}|)\right\} - \frac{1}{\sigma_e^2}\mathbb{C}\text{ov}_{x,\varepsilon}(z_{kj}, \varepsilon_{kj}) = 0, \tag{24}$$

where $\mathbb{E}_{x,\varepsilon}$ *and* $\mathbb{C}\text{ov}_{x,\varepsilon}$ *denote expectation and covariance conditionally on* $(\mathbf{x}, \varepsilon)$.

Proof Denote the left-hand side of Eq. (24) by Δ. By assumption, the response $y_{kj}^{(\lambda_0)}$ conditionally on v_k and $\mathbf{x}_{kj}$ follows $N(\mathbf{x}_{kj}^\top\beta + v_k, \sigma_e^2)$. Since $\mathbb{E}(\varepsilon_{kj}) = 0$ and $z_{kj} = y_{rj}^{(\lambda_0)}\log(|y_{kj}|) = \lambda_0^{-1}y_{rj}^{(\lambda_0)}\log(|y_{kj}^{(\lambda_0)}|)$, it follows that

$$\Delta = \frac{1}{\lambda_0} + \frac{1}{\lambda_0}\int_{-\infty}^{\infty}\log(|t|)\phi\left(\frac{t-a}{\sigma_e}\right)\frac{dt}{\sigma_e} - \frac{1}{\lambda_0}\int_{-\infty}^{\infty}t(t-a)\log(|t|)\phi\left(\frac{t-a}{\sigma_e}\right)\frac{dt}{\sigma_e^3},$$

where we denote $a = \mathbf{x}_{kj}^\top\beta + v_k$ for short. By transforming $u = t/\sigma_e$ and $b = a/\sigma_e$, we further have

$$\begin{aligned}\Delta &= \frac{1}{\lambda_0} + \frac{1}{\lambda_0}\int_{-\infty}^{\infty}\log(|u|\sigma_e)\phi(u-b)du - \frac{1}{\lambda_0}\int_{-\infty}^{\infty}u(u-b)\log(|u|\sigma_e)\phi(u-b)du \\ &= \frac{1}{\lambda_0} + \frac{1}{\lambda_0}\int_{-\infty}^{\infty}\{1-u(u-b)\}\log(|u|)\phi(u-b)du.\end{aligned}$$

Using $d\{\phi(u-b)u\} = \{1-u(u-b)\}\phi(u)du$ and integration by parts, we have

$$\begin{aligned}\Delta &= \frac{1}{\lambda_0} + \frac{1}{\lambda_0}\int_{-\infty}^{\infty} \log(|u|)d\{\phi(u-b)u\} \\ &= \frac{1}{\lambda_0} + \frac{1}{\lambda_0}\log(|u|)\phi(u-b)u\Big|_{-\infty}^{\infty} - \frac{1}{\lambda_0}\int_{-\infty}^{\infty}\phi(u-b)du \\ &= 0.\end{aligned}$$

□

References

Battese GE, Harter RM, Fuller WA (1988) An error components model for prediction of county crop area using survey and satellite data. J Am Stat Assoc 83:28–36

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc Ser B 26:211–252

Chen J, Variyath AM, Abraham B (2008) Adjusted empirical likelihood and its properties. J Comput Gr Stat 17:426–443

Duan N (1983) Smearing estimate: a nonparametric retransformation method. J Am Stat Assoc 78:605–610

Datta GS, Lahiri P (2000) A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Stat Sin 10:613–627

Fay RE, Herriot RA (1979) Estimates of income for small places: an application of James-Stein procedures to census data. J Am Stat Assoc 74:269–277

González-Manteiga W, Lombardía MJ, Molina I, Morales D, Santamaría L (2008) Bootstrap mean squared error of a small-area EBLUP. J Stat Comput Simul 78:443–462

Gurka MJ, Edward LJ, Muller KE, Kupper LL (2006) Extending the Box-Cox transformation to the linear mixed model. J R Stat Soc Ser A 169:273–288

Hall P, Maiti T (2006) Nonparametric estimation of mean-squared prediction error in nested-error regression models. Ann Stat 34:1733–1750

Jiang J, Nguyen T (2012) Small area estimation via heteroscedastic nested-error regression. Can J Stat 40:588–603

John JA, Draper NR (1980) An alternative family of transformations. Appl Stat 29:190–197

Owen AB (1990) Empirical likelihood ratio confidence regions. Ann Stat 18:90–120

Owen AB (2001) Empirical likelihood. Chapman and Hall/CRC, New York

Pfeffermann D (2013) New important developments in small area estimation. Stat Sci 28:40–68

Pfeffermann D, Correa S (2012) Emprical bootstrap bias correction of prediction mean square error in small area estimation. Brometrika 99:457–472

Prasad NGN, Rao JNK (1990) The estimation of the mean squared error of small-area estimators. J Am Stat Assoc 85:163–171

Qin J, Lawless J (1994) Empirical likelihood and general estimating equations. Ann Stat 22:300–325

Rao JNK (2003) Small area estimation. Wiley, Hoboken

Rao JNK, Molina I (2015) Small area estimation, 2nd edn. Wiley, Hoboken

Schmid T, Munnich RT (2014) Spatial robust small area estimation. Stat Pap 55:653–670

Sinha SK, Rao JNK (2009) Robust estimation of small area estimation. Can J Stat 37:381–399

Sugasawa S, Kubokawa T (2014) Estimation and prediction in transformed nested error regression models. Manuscript. arXiv:1410.8269v1

Sugasawa S, Kubokawa T (2015) Box-Cox transformed linear mixed models for positive-valued and clustered data. Manuscript. arXiv:1502.03193v2

Yang ZL (2006) A modified family of power transformations. Econ Lett 92:14–19