Biometrical Journal **52** (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

22!

Empirical Bayes small area prediction under a zero-inflated lognormal model with correlated random area effects

Xiaodan Lyu [*,1], **Emily J. Berg** [1], and **Heike Hofmann** [1]

[1] Department of Statistics, Iowa State University, Ames, Iowa, 50010, U.S.A.

Many variables of interest in agricultural or economical surveys have skewed distributions and can equal zero. Our data are measures of sheet and rill erosion called RUSLE2. Small area estimates of mean RUSLE2 erosion are of interest. We use a zero-inflated lognormal mixed effects model for small area estimation. The model combines a unit-level lognormal model for the positive RUSLE2 responses with a unit-level logistic mixed effects model for the binary indicator that the response is nonzero. In the CEAP data, counties with a higher probability of nonzero responses also tend to have a higher mean among the positive RUSLE2 values. We capture this property of the data through an assumption that the pair of random effects for a county are correlated. We develop empirical Bayes small area predictors and a bootstrap estimator of the MSE. In simulations, the proposed predictor is superior to simpler alternatives. We then apply the method to construct empirical Bayes predictors of mean RUSLE2 erosion for South Dakota counties. To obtain auxiliary variables for the population of cropland in South Dakota, we integrate a satellite derived land cover map with a geographic database of soil properties. We provide an R Shiny application called `viscover` (available at `https://lyux.shinyapps.io/viscover/`) to visualize the overlay operations required to construct the covariates. On the basis of bootstrap estimates of the mean square error, we conclude that the empirical Bayes predictors of mean RUSLE2 erosion are superior to direct estimators.

Key words: Correlated random effects; Empirical Bayes; Sheet and rill erosion; Small area prediction; Zero-inflated lognormal

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1022/bimj.XXXXXXX

1 Introduction

In small area estimation, direct estimators for domains of interest are considered unreliable due to small sample sizes. Model-based small area estimators (Rao and Molina, 2015) attain efficiency gains relative to design-based estimators through the use of auxiliary information, often assumed known for the full population. Early and influential approaches to small area estimation use linear mixed effects models with symmetric error distributions (Battese et al., 1988; Fay III and Herriot, 1979).

We consider small area estimation for skewed response variables that are contaminated with zeros. Examples of variables that exhibit these distributional properties include household expenditures on durable goods (Tobin, 1958), agricultural production (Dreassi et al., 2014; Karlberg, 2015), and expenditures for medical care. Min and Agresti (2002) provide additional examples in the context of a review of regression models for semi-continuous data. When distributions are skewed and inflated with zeros, the assumptions of the linear mixed effects model may not hold.

*Corresponding author: e-mail: annielyu8@gmail.com, homepage: https://annielyu.com/

1.1 Motivating CEAP RUSLE2 data

The data that motivate our study are from the Conservation Effects Assessment Project (CEAP), a survey conducted from 2003-2006 through a cooperative agreement between the Natural Resources Conservation Service of the United States Department of Agriculture and Iowa State University. The response variables in CEAP are several measures of soil and nutrient loss on crop fields due to different types of water and wind erosion. We focus on a particular measure of sheet and rill erosion obtained by processing survey responses collected in CEAP through a computer model, the Revised Universal Soil Loss Equation – 2 (RUSLE2).

Figure 1 demonstrates the challenges associated with specifying an appropriate small area model for the CEAP RUSLE2 data. The histogram in the left panel of Figure 1 shows that the distribution of RUSLE2 is highly skewed right, and approximately 15% of the RUSLE2 measurements are equal to zero. The scatterplot in the right panel of Figure 1 shows that counties with higher mean values of positive erosion (in the log scale) are less likely to have observed zeros. An appropriate small area model for the CEAP RUSLE2 data will reflect both of these characteristics: the right skew in the distribution of positive RUSLE2 values and the positive correlation between the county mean of log positive erosion and the county level probability that erosion is nonzero.

1.2 Zero-inflated lognormal model with correlated area random effects

We define a zero-inflated lognormal model with correlated random effects to address the challenges in modeling the CEAP RUSLE2 data. Let $i = 1, \ldots, D$ index the small domains (South Dakota counties, for the CEAP application) of interest and $j = 1, \ldots, N_i$ index the units in the population for area i. We assume that a probability sample is selected and let s_i be the set of sampled units for area i with $|s_i| = n_i$. The observed response variable (RUSLE2 in the CEAP application), denoted as y^*_{ij}, satisfies

$$y^*_{ij} = \delta_{ij} y_{ij}, \tag{1}$$

where δ_{ij} is a Bernoulli random variable with probability p_{ij} of being 1, and $y_{ij} > 0$. The quantities of interest are the area means defined by $\bar{y}^*_{N_i} = N_i^{-1} \sum_{j=1}^{N_i} y^*_{ij}$ for $i = 1, \ldots, D$. In the model, defined precisely below, we assume y_{ij} satisfies a lognormal mixed effects model and relate p_{ij} to the vector of covariates using a general link function $g(\cdot)$. Assume

$$\log(y_{ij}) = \beta_0 + \boldsymbol{z}'_{1,ij}\boldsymbol{\beta}_1 + u_i + e_{ij}, \tag{2}$$

and

$$g(p_{ij}) = \alpha_0 + \boldsymbol{z}'_{2,ij}\boldsymbol{\alpha}_1 + b_i, \tag{3}$$

where $g(\cdot) : (0,1) \to (-\infty, \infty)$ is a parametric, one-to-one link function, $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$, and the pair of random area effects (u_i, b_i) satisfies

$$\begin{pmatrix} u_i \\ b_i \end{pmatrix} \overset{iid}{\sim} \text{BVN}(\mathbf{0}, \boldsymbol{\Sigma}_{ub}), \boldsymbol{\Sigma}_{ub} = \begin{pmatrix} \sigma_u^2 & \sigma_{ub} \\ \sigma_{ub} & \sigma_b^2 \end{pmatrix},$$

where $\sigma_{ub} = \rho\sigma_u\sigma_b$. The pairs of random variables $\{(u_i, b_i) : i = 1, \ldots, D\}$ are mutually independent and are also independent of e_{ij} for $i = 1, \ldots, D$ and $j = 1, \ldots, N_i$. Implications of this dependence structure are that $\delta_{ij} \perp u_i \mid b_i$, and $\delta_{ij} \perp e_{ij} \mid (b_i, u_i)$, where $\perp$ denotes independence. When the link function $g(\mu)$ is chosen to be the logit link $\log(\mu/(1-\mu))$, model (3) is called logistic mixed effects model.

The model (2-3) represents the main features of the data seen in Figure 1. The lognormal model (2) captures the skewed distribution observed in the left panel of Figure 1. The binary part (3) accounts for the nontrivial proportion of zeros in the CEAP data set. The important parameter $\rho \in (-1, 1)$, the correlation

between b_i and u_i, enables us to represent the correlation between the positive and binary parts observed in the right panel of Figure 1 through our model. In a preliminary analysis of the CEAP data, we fit a simplified version of the model (2-3) with $\rho = 0$. In this "independence model", the lognormal mixed effects model for the positive RUSLE2 values is independent of the logistic mixed effects model for the probability of a nonzero. The sample correlation between random effects from these two independent models for the positive and binary parts suggests that the population correlation is nontrivial. Our model development in this paper allows us to investigate the significance of the correlation between b_i and u_i formally through a bootstrap interval.

1.3 Related small area procedures for skewed, binary, or zero-inflated data

The model (2-3) advances existing small area literature primarily through the introduction of the correlation parameter ρ in the lognormal framework and secondarily through the development of small area predictors and mean squared error (MSE) estimators in Section 2. We review the literature on which our model and procedure build in this subsection. We first review procedures that address either skewed positive data or binary data individually. We then turn attention to models for zero-inflated data.

Several small area models have been developed for positive, skewed data with support $(0, \infty)$. Most relevant for our work, Berg and Chandra (2014) develop closed form expressions for the empirical Bayes (EB) small area predictor and corresponding mean squared error (MSE) estimator under the assumptions of a lognormal mixed effects model. They apply the mixed effects small area model to construct predictors for a subset of the CEAP data collected in Iowa, a highly cultivated and relatively homogeneous state with a negligible number of observed zeros. For states with greater heterogeneity than Iowa with respect to agriculture, appropriately modeling observed zeros is an important issue for small area estimation. Several studies extend Berg and Chandra (2014). Molina et al. (2018) study MSE estimation for the lognormal small area model of Berg and Chandra (2014) more rigorously, and Zimmermann and Münnich (2018) consider informative sampling. Molina and Rao (2010) and Marhuenda et al. (2017) generalize the lognormal small area model to an arbitrary class of transformations. We considered the Box-Cox family of transformations for the positive component as well as parametrized families of link functions for the binary component. We focus on the log transformation with the logit link function because those transformations fit the CEAP data well, as we demonstrate in Section 4. Distributions other than the lognormal distribution for skewed, positive data have also been considered for small area estimation. Berg et al. (2016) compare properties of small area predictors constructed under a mixed effects lognormal model to those based on a generalized linear mixed model (GLMM) with a gamma response distribution in a simulation study. Jiang (2003) develops EB small area predictors under the assumptions of a general GLMM.

A vast literature also exists on small area estimation for binary data. In the context of a unit-level logistic mixed effects model, González-Manteiga et al. (2007) investigate a bootstrap MSE estimator for a small area predictor of a binary response variable constructed using the penalized quasi-likelihood method of Schall (1991). Hobza and Morales (2016) improve upon González-Manteiga et al. (2007) by using a Laplace approximation for the maximum likelihood estimator and EB prediction. Jiang and Lahiri (2001) develops a rigorous theory for EB prediction under the unit-level logistic mixed effects model, where the parameters are estimated using simulated method of moments. Extensions of the logistic mixed effects model with a single normally distributed random effect have been developed to incorporate a semiparametric model for the random effects (Marino et al., 2019) and temporal data (Hobza et al., 2018).

Pfeffermann et al. (2008) combine the logistic mixed effects model with a unit-level linear mixed effects model to develop a unit-level small area model for zero-inflated data. In constructing small area estimates of literacy rates, Pfeffermann et al. (2008) use this two-part random effects model to accommodate a large number of observed literacy scores of zero. A logistic mixed effects model is used to describe the probability of a nonzero, and a linear mixed model is used for the nonzero literacy scores. Pfeffermann et al. (2008) use Bayesian methods for inference. Our development addresses many of the challenges associated with a frequentist analysis of a two-part small area model raised in Pfeffermann et al. (2008).

The use of a linear model for the positive component, as in Pfeffermann et al. (2008), is undesirable for CEAP because the RUSLE2 data are highly skewed and have nonlinear associations with covariates.

Existing approaches to small area estimation for zero-inflated skewed data include Dreassi et al. (2014) and Chandra and Chambers (2016). Dreassi et al. (2014) develop a zero-inflated gamma model and apply the zero-inflated model to estimate grape wine production in small areas in Italy. Similar to Pfeffermann et al. (2008), Dreassi et al. (2014) use Bayesian methods for inference. Chandra and Chambers (2016) extend the lognormal model of Berg and Chandra (2014) to a two-part model, where the binary component is independent of the positive, lognormal component. Chandra and Chambers (2016) use penalized quasi-likelihood (PQL) to define a "plug-in" type of predictor for the binary component, similar to González-Manteiga et al. (2007). As discussed in Hobza and Morales (2016), the PQL predictor is not an EB predictor.

When developing a two-part zero-inflated mixed effects model, one issue to consider is the correlation between the random effects in the positive component and the random effects in the binary component. In the model (2-3), this correlation is represented through the parameter ρ. The Bayesian methods of Dreassi et al. (2014) and Pfeffermann et al. (2008) are flexible enough to allow for a nonzero correlation. Implicit in the "plug-in" approach of Chandra and Chambers (2016) is an assumption that the model for the positive component is independent of the model for the binary component. Frequentist estimation and inference for the two-part model with correlated random effects creates new analytical challenges. Nonetheless, we prefer to adopt a frequentist approach since (1) we lack prior information that would guide an appropriate choice of prior distributions and (2) constructing a posterior distribution for all elements of the large population of cropland in South Dakota is computationally expensive. We address the analytical challenges arising in the frequentist analysis of the zero-inflated lognormal model through our development of a maximum likelihood estimator and an EB predictor in Section 2.

1.4 Auxiliary information acquisition for small area models

To construct EB predictors, we require the covariates to be available for every element of the population. The requirement that the covariates are known for all population elements results from the nonlinear model form. For a linear model (Battese et al., 1988), in contrast, the population means of covariates for each area are sufficient. For nonlinear small area models, the assumption that the covariate is available for all population elements is common. One of the most labor-intensive steps in small area estimation in practice is obtaining the auxiliary variables for the full population. Many applications in small area estimation touch upon that issue lightly (i.e., Pfeffermann et al. (2008)) or resort to incomplete auxiliary information, which inflates the MSE of the predictor (Erciulescu and Fuller, 2016).

We provide a thorough treatment of the issue of obtaining population level auxiliary information. We integrate a land cover map derived from satellite imagery with administrative data on soil properties to obtain covariates that are known for the full population and relate to several factors thought to impact erosion. The covariates $\boldsymbol{z}_{1,ij}$ and $\boldsymbol{z}_{2,ij}$ for the positive and binary parts in the model (2-3) can be the same or different. Besides subject matter and data availability, the choice of covariates for each model component also depends on appropriate variable selection techniques. We use a combination of the science underlying erosion and statistical model comparison techniques to select the acquired covariates for the CEAP data analysis in Section 4.

1.5 Outline of our approach

We develop EB predictors of small area means and corresponding MSE estimators based on the zero-inflated lognormal model defined in (2-3). The EB predictor has theoretical support because it is an estimator of the minimum MSE unbiased predictor. The EB predictor enables a computationally simple estimator of the leading term in the MSE as well as parametric bootstrap MSE estimators. We present

the small area predictors based on the zero-inflated lognormal model with correlated random area effects in Section 2. In Section 3, we compare our proposed EB predictor to alternatives through simulations. In Section 4, we describe how we obtain the auxiliary variables for the full population, apply the method to the CEAP RUSLE2 data collected in South Dakota, and give predictions of county-level average soil erosion with associated standard errors. We develop a visualization tool that demonstrates the overlay operations needed to construct the covariates and also helps verify that the overlay operation works appropriately. The tool is featured in the RStudio Shiny gallery and available online at `https://lyux.shinyapps.io/viscover/`. Concluding remarks are given in Section 5. The Appendix provides the computational details for maximum likelihood estimation including the functional form of the score equation. Tables and figures not required to communicate the main ideas are deferred to Supporting Information available on the journal's web page (http://onlinelibrary.wiley.com/doi/10.1002/bimj.202000029/suppinfo) We develop a `R` (R Core Team, 2019) package `saezero` (Lyu, 2020) to implement the maximum likelihood estimation and empirical Bayes prediction methods described in this paper.

2 Empirical Bayes prediction for the zero-inflated lognormal model

To define empirical Bayes predictors and MSE estimators under model (1), we introduce additional notation. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}', \sigma_u^2, \sigma_b^2, \sigma_e^2, \rho)'$ denote the vector of fixed model parameters to be estimated, where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1')'$ and $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}_1')'$. As we illustrate through the data analysis, one can add parameters of parametrized families of link functions or transformations.

Assume a sample is selected, and denote the index sets for sampled and nonsampled elements in area i by s_i and $\bar{s}_i$, respectively. Assume that y_{ij}^* is observed for all sampled elements. As discussed in Section 1.3, assume that the covariate $\boldsymbol{z}_{ij} = (\boldsymbol{z}_{1,ij}', \boldsymbol{z}_{2,ij}')'$ is observed for every element in the population, where $\boldsymbol{z}_{1,ij}$ and $\boldsymbol{z}_{2,ij}$, respectively, are the covariates in the models (2) for the positive part and in the model (3) for the binary part. Denote the observed data as $(\boldsymbol{y}^*, \boldsymbol{z}) = \{y_{ij}^*, j \in s_i, i = 1, \ldots, D\} \cup \{\boldsymbol{z}_{ij} : j = 1, \ldots, N_i; i = 1, \ldots, D\}$. As discussed in Section 1.4, the requirement that the covariate $\boldsymbol{z}_{ij}$ is known for every element of the population results from the nonlinear model form. We explain how we achieve this requirement for the CEAP data analysis in Section 4.

2.1 Minimum mean square error predictor of small area mean

A common objective function for defining a predictor in small area estimation is squared error loss. It is straightforward to show that the optimal predictor of $\bar{y}_{N_i}^*$ under squared error loss is $\mathrm{E}\{\bar{y}_{N_i}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\}$, the conditional expectation of the population mean given the observed data. We call $\mathrm{E}\{\bar{y}_{N_i}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\}$ the minimum mean square error (MMSE) predictor. A synonym for the MMSE predictor used, for example, in Molina and Rao (2010) is the term "best predictor" (BP). The MMSE predictor is also often called the "Bayes" predictor. To stress that the MMSE predictor of $\bar{y}_{N_i}^*$ depends on the unknown $\boldsymbol{\theta}$, we denote the MMSE predictor by $\hat{\bar{y}}_{N_i}^{*\mathrm{MMSE}}(\boldsymbol{\theta})$.

In this section, we derive a form for the conditional expectation defining the MMSE predictor for the specific model (2-3). The MMSE predictor is difficult to attain because the required conditional expectation involves a bivariate integral over the distribution $f(u_i, b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))$. We present a form for the MMSE predictor as a univariate integral in Theorem 2.1 below. The key idea of the derivation is that the conditional distribution of u_i given b_i and $(\boldsymbol{y}^*, \boldsymbol{z})$ has a standard form. This allows us to transform the bivariate integral defining the MMSE predictor to a univariate integral.

Theorem 2.1 *The minimum MSE predictor of $\bar{y}_{N_i}^*$ under model (2-3) has the form*

$$\hat{\bar{y}}_{N_i}^{*\mathrm{MMSE}}(\boldsymbol{\theta}) = E\{\bar{y}_{N_i}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\} = \frac{1}{N_i}\Big[\sum_{j \in s_i} y_{ij}^* + \sum_{j \in \bar{s}_i} E\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}\Big], \tag{4}$$

where

$$E\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} = h_{ij}(\boldsymbol{\theta}) \frac{\int \pi_{s_i}(b_i) \exp\left\{-\frac{1}{2v_i}(b_i - m_i)^2\right\} p_{ij}(b_i) \eta(b_i)\, \mathrm{d}b_i}{\int \pi_{s_i}(b_i) \exp\left\{-\frac{1}{2v_i}(b_i - m_i)^2\right\} \mathrm{d}b_i}, \tag{5}$$

$$(m_i, v_i) = (\bar{\tilde{r}}_i \rho \sigma_u \sigma_b (\sigma_u^2 + \sigma_e^2/\tilde{n}_i)^{-1}, \sigma_b^2\{(1-\rho^2)\sigma_u^2 + \sigma_e^2/\tilde{n}_i\}(\sigma_u^2 + \sigma_e^2/\tilde{n}_i)^{-1}),$$

$h_{ij}(\boldsymbol{\theta}) = \exp(\beta_0 + \boldsymbol{z}'_{1,ij}\boldsymbol{\beta}_1 + \gamma_i \bar{\tilde{r}}_i + \gamma_i \sigma_e^2 \tilde{n}_i^{-1}/2 + \sigma_e^2/2)$, $\pi_{s_i}(b_i) = \prod_{j \in s_i} \left[p_{ij}(b_i)^{\delta_{ij}} \{1 - p_{ij}(b_i)\}^{(1-\delta_{ij})}\right]$, $p_{ij}(b_i) = g^{-1}(\alpha_0 + \boldsymbol{z}'_{2,ij}\boldsymbol{\alpha}_1 + b_i)$, $\eta(b_i) = \exp\{(1-\gamma_i)\rho\sigma_u/\sigma_b b_i\}$, $\gamma_i = (1-\rho^2)\sigma_u^2\{(1-\rho^2)\sigma_u^2 + \sigma_e^2/\tilde{n}_i\}^{-1}$, $\bar{\tilde{r}}_i = \tilde{n}_i^{-1} \sum_{j \in s_i} \tilde{r}_{ij}$, $\tilde{r}_{ij} = \delta_{ij}\{\log(y_{ij}) - \beta_0 - \boldsymbol{z}'_{1,ij}\boldsymbol{\beta}_1\}$, *and* $\tilde{n}_i = \sum_{j \in s_i} \delta_{ij}$.

Proof. The main steps of the proof are to first find a form for $f(u_i \mid b_i, (\boldsymbol{y}^*, \boldsymbol{z}))$ and to second find a form for $f(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))$. By standard properties of bivariate normal distributions, $u_i \mid b_i \sim N(\rho\sigma_u\sigma_b^{-1}b_i, (1-\rho^2)\sigma_u^2)$. Additionally, recall that the assumptions of model (2-3) imply that $\delta_{ij} \perp e_{ij} \mid b_i$ and $u_i \perp \delta_{ij} \mid b_i$. This allows us to specify the model for $\log(y_{ij})$ as a unit-level model in new parameters. Specifically,

$$\log(y_{ij}) = \beta_0 + \boldsymbol{z}'_{1,ij}\boldsymbol{\beta}_1 + \rho\sigma_u\sigma_b^{-1}b_i + k_i + e_{ij},$$

where $k_i = u_i - \rho\sigma_u\sigma_b^{-1}b_i$, $k_i \sim N(0, (1-\rho^2)\sigma_u^2)$, and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$. Then standard properties of the linear mixed effects model (i.e., Battese et al. (1988)) imply that $u_i \mid \{b_i, (\boldsymbol{y}^*, \boldsymbol{z})\} \sim N(\tilde{\mu}_{u_i}, \tilde{\sigma}_{u_i}^2)$, where $\tilde{\mu}_{u_i} = \gamma_i \bar{\tilde{r}}_i + (1-\gamma_i)\rho\sigma_u\sigma_b^{-1}b_i$ and $\tilde{\sigma}_{u_i}^2 = \gamma_i\sigma_e^2/\tilde{n}_i$. When $\tilde{n}_i = 0$ which indicates there is no positive response in area i, it holds that $\gamma_i = 0, \tilde{\mu}_{u_i} = \rho\sigma_u\sigma_b^{-1}b_i$ and $\tilde{\sigma}_{u_i}^2 = (1-\rho^2)\sigma_u^2$. To derive the density of the conditional distribution of b_i given the observed data, note that $f(u_i \mid b_i, (\boldsymbol{y}^*, \boldsymbol{z}))f(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})) = f(u_i, b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})) \propto f(\boldsymbol{y}_i^* \mid u_i, b_i, \boldsymbol{z})f(u_i \mid b_i, \boldsymbol{z})f(b_i \mid \boldsymbol{z})$ implies

$$f(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})) \propto \frac{f(\boldsymbol{y}_i^* \mid u_i, b_i, \boldsymbol{z})f(u_i \mid b_i, \boldsymbol{z})f(b_i, \boldsymbol{z})}{f(u_i \mid b_i, (\boldsymbol{y}^*, \boldsymbol{z}))}, \tag{6}$$

where $\boldsymbol{y}_i^* = (y_{i1}^*, \ldots, y_{in_i}^*)'$. Plugging in the known terms on the right side of (6) gives $f(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})) \propto \pi_{s_i}(b_i)\exp\{-(b_i - m_i)^2/(2v_i)\}$. When $\tilde{n}_i = 0$ or $\rho = 0$, it can be shown that $m_i = 0$ and $v_i = \sigma_b^2$. Let $\phi(\cdot)$ denote the probability density function of a standard normal distribution. Then, the conditional probability density function of b_i given the observed data is

$$h(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})) = \frac{\frac{1}{\sqrt{v_i}}\pi_{s_i}(b_i)\phi(\frac{b_i - m_i}{\sqrt{v_i}})}{\int \frac{1}{\sqrt{v_i}}\pi_{s_i}(b_i)\phi(\frac{b_i - m_i}{\sqrt{v_i}})\, \mathrm{d}b_i}.$$

We are now able to express the conditional expectation of y_{ij}^* as a univariate integral. By definition,

$$\begin{aligned}
\mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} &= \mathrm{E}\{\delta_{ij} \exp(\beta_0 + \boldsymbol{z}'_{1,ij}\boldsymbol{\beta}_1 + u_i + e_{ij}) \mid (\boldsymbol{y}^*, \boldsymbol{z})\} \\
&= \exp(\beta_0 + \boldsymbol{z}'_{1,ij}\boldsymbol{\beta}_1 + \sigma_e^2/2)\mathrm{E}\{p_{ij}(b_i)\exp(u_i) \mid (\boldsymbol{y}^*, \boldsymbol{z})\} \\
&= \exp(\beta_0 + \boldsymbol{z}'_{1,ij}\boldsymbol{\beta}_1 + \sigma_e^2/2)\mathrm{E}[\mathrm{E}\{p_{ij}(b_i)\exp(u_i) \mid b_i, (\boldsymbol{y}^*, \boldsymbol{z})\} \mid (\boldsymbol{y}^*, \boldsymbol{z})] \\
&= h_{ij}(\boldsymbol{\theta})\mathrm{E}\{p_{ij}(b_i)\eta(b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z})\},
\end{aligned}$$

where the second equal sign results from the independence of e_{ij} and e_{ik} for $j \neq k$, and the final equal sign results from the conditional normal distribution of u_i given b_i and the observed data as well as the moment generating function of a normal distribution. The expression for the conditional expectation as a univariate integral in the statement of Theorem 2.1 now follows. □

Observe that we derive a minimum MSE predictor for y_{ij}^* in the original scale of the observed responses, not in the log scale. By double conditioning, $\mathrm{E}[\mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} - y_{ij}^*] = \mathrm{E}(\mathrm{E}[\mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} - y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})]) = 0$. A correction for the transformation between $\log(y_{ij})$ and y_{ij} is implicit in the development of the minimum MSE predictor.

2.2 Empirical Bayes predictor of small area mean

The minimum MSE predictor is defined through the integral in (5) and depends on the unknown $\boldsymbol{\theta}$. An empirical Bayes (EB) predictor is an estimate of the MMSE predictor (Rao and Molina, 2015, p. 271). Calculating an EB predictor in practice requires estimating the parameter vector $\boldsymbol{\theta}$ and approximating the integral in (5). We use maximum likelihood to estimate $\boldsymbol{\theta}$ and approximate the integral in (5) using Gauss-Hermite quadrature (Golub and Welsch, 1969).

An advantage of converting bivariate integrals with respect to the conditional distribution of $(u_i, b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z})$ to univariate integrals with respect to the distribution of $b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})$ is that one can easily approximate an integral with respect to the distribution $h(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))$ using a Gauss-Hermite approximation (Smyth, 2014). For an arbitrary function $q(b_i)$, a Gauss-Hermite approximation to $\mathrm{E}\{q(b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z})\}$ is given by

$$\mathrm{E}_A\{q(b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} = \frac{\sum_{k=1}^K \pi_{s_i}(b_{i,k}) q(b_{i,k}) w_k(m_i, v_i)}{\sum_{k=1}^K \pi_{s_i}(b_{i,k}) w_k(m_i, v_i)}, \tag{7}$$

where $b_{i,k}$ for $k = 1, \ldots, K$ are specified nodes, and $w_k(m_i, v_i)$ is a weight defined to approximate expectation with respect to a normal distribution with mean m_i and variance v_i. We obtain the nodes and weights using the `R` function `gauss.quad.prob` in the `R` package `statmod` (Giner and Smyth, 2016). We use the approximation (7) to define maximum likelihood estimators and to construct EB predictors.

Using the same reasoning used in (6) of the derivation of $h(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))$, we have $f(\boldsymbol{y}_i^* \mid \boldsymbol{\theta}) = [f(u_i \mid b_i, (\boldsymbol{y}^*, \boldsymbol{z})) f(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))]^{-1} f(\boldsymbol{y}_i^* \mid u_i, b_i) f(u_i \mid b_i) f(b_i)$. The terms on the right side of this expression have known forms from the proof of Theorem 2.1. Substitution of these known forms gives the likelihood function $L(\boldsymbol{\theta}) = \prod_{i=1}^D L_i(\boldsymbol{\theta})$ where

$$\begin{aligned} L_i(\boldsymbol{\theta}) &= \frac{\prod_{j \in s_i} [\frac{1}{\sigma_e} \phi(\frac{r_{ij} - u_i}{\sigma_e})]^{\delta_{ij}} \frac{1}{\sqrt{(1-\rho^2)\sigma_u^2}} \phi(\frac{u_i - \rho\sigma_u \sigma_b^{-1} b_i}{\sqrt{(1-\rho^2)\sigma_u^2}}) \pi_{s_i}(b_i) \frac{1}{\sigma_b} \phi(\frac{b_i}{\sigma_b})}{\frac{1}{\tilde{\sigma}_{u_i}} \phi(\frac{u_i - \tilde{\mu}_{u_i}}{\tilde{\sigma}_{u_i}}) \pi_{s_i}(b_i) \frac{1}{\sqrt{v_i}} \phi(\frac{b_i - m_i}{\sqrt{v_i}}) / \int \pi_{s_i}(b_i) \frac{1}{\sqrt{v_i}} \phi(\frac{b_i - m_i}{\sqrt{v_i}}) \mathrm{d}b_i} \\ &= \frac{(1-\gamma_i)^{1/2} (v_i/\sigma_b^2)^{1/2}}{(2\pi\sigma_e^2)^{\tilde{n}_i/2}} \exp\left(\frac{\gamma_i \bar{\tilde{r}}_i^2}{2\sigma_e^2/\tilde{n}_i} + \frac{m_i^2}{2v_i} - \frac{\sum_j \tilde{r}_{ij}^2}{2\sigma_e^2} \right) \int \pi_{s_i}(b_i) \frac{1}{\sqrt{v_i}} \phi(\frac{b_i - m_i}{\sqrt{v_i}}) \mathrm{d}b_i. \end{aligned}$$

A Gauss-Hermite approximation to the log likelihood is then $\ell_A(\boldsymbol{\theta}) = \sum_{i=1}^D \log(L_{i,A}(\boldsymbol{\theta}))$, where $L_{i,A}(\boldsymbol{\theta})$ approximates $L_i(\boldsymbol{\theta})$ by substituting $\int \pi_{s_i}(b_i) {v_i}^{-1/2} \phi((b_i - m_i) {v_i}^{-1/2}) \mathrm{d}b_i$ with $\sum_{k=1}^K \pi_{s_i}(b_{i,k}) w_k(m_i, v_i)$. We define the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ to satisfy $\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} \ell_A(\boldsymbol{\theta})$. To maximize $\ell_A(\boldsymbol{\theta})$ operationally, we use the `R` function `optim` in the `R` package `stats` (R Core Team, 2019), as explained in more detail in the Appendix. To calculate the empirical Bayes predictor, we replace the unknown $\boldsymbol{\theta}$ with an estimator and approximate expectations with respect to the distribution $h(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))$ using the Gauss-Hermite approximation (7). We estimate (5), the conditional expectation of a nonsampled y_{ij}^* given the observed data, by

$$\hat{y}_{ij}^{*\mathrm{MMSE}}(\hat{\boldsymbol{\theta}}) = h_{ij}(\hat{\boldsymbol{\theta}}) \mathrm{E}_A\{\hat{p}_{ij}(b_i) \hat{\eta}(b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\},$$

where $\hat{p}_{ij}(b_i) = g^{-1}(\hat{\alpha}_0 + \boldsymbol{z}_{2,ij}' \hat{\boldsymbol{\alpha}}_1 + b_i)$ and $\hat{\eta}(b_i) = \exp\{(1 - \hat{\gamma}_i) \hat{\rho} \hat{\sigma}_u \hat{\sigma}_b^{-1} b_i\}$. The EB predictor is then defined by

$$\hat{\bar{y}}_{N_i}^{*\mathrm{EB}} = \frac{1}{N_i} \Big\{ \sum_{j \in s_i} y_{ij}^* + \sum_{j \in \bar{s}_i} \hat{y}_{ij}^{*\mathrm{MMSE}}(\hat{\boldsymbol{\theta}}) \Big\}. \tag{8}$$

The term in the EB predictor that provides the predictor of p_{ij} has connections to EB predictors of proportions developed under the unit-level logistic mixed effects model in Hobza and Morales (2016) and Hobza et al. (2018).

2.3 MSE of the EB predictor

Theorem 2.2 gives an expression for the MSE of the EB predictor (8) as the sum of the MSE of the MMSE predictor (4) and a second term that accounts for the variance of $\hat{\boldsymbol{\theta}}$. We express the MSE of the MMSE predictor as a univariate integral, which we approximate through Gauss-Hermite quadrature after substituting the unknown parameters with estimators. We estimate the second term in the MSE using the parametric bootstrap, a resampling procedure that involves simulating repeatedly from the estimated model (Rao and Molina, 2015, p. 226).

Theorem 2.2 *Under model (2-3),*

$$\mathrm{MSE}(\hat{\bar{y}}_{N_i}^{*\mathrm{EB}}) = E\{(\hat{\bar{y}}_{N_i}^{*\mathrm{EB}} - \bar{y}_{N_i}^{*})^2\} = M_{1i}(\boldsymbol{\theta}) + M_{2i}(\boldsymbol{\theta}),$$

where

$$M_{1i}(\boldsymbol{\theta}) = E\left(\frac{1}{N_i^2}\sum_{j\in\bar{s}_i}\sum_{k\in\bar{s}_i}\Big[E\{y_{ij}^{*}y_{ik}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\} - E\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\}E\{y_{ik}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\}\Big]\right),$$

$$E\{y_{ij}^{*}y_{ik}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\} = \begin{cases} h_{ij}h_{ik}\exp(\tilde{\sigma}_{u_i}^2)E\{p_{ij}p_{ik}\eta(2b_i) \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\} & \text{if } j \neq k \\ h_{ij}^2\exp(\tilde{\sigma}_{u_i}^2+\sigma_e^2)E\{p_{ij}\eta(2b_i) \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\} & \text{if } j = k \end{cases},$$

and $M_{2i}(\boldsymbol{\theta}) = E\{(\hat{\bar{y}}_{N_i}^{*\mathrm{EB}} - \hat{\bar{y}}_{N_i}^{*\mathrm{MMSE}})^2\}$.

Proof. Expanding the square defining the MSE, we have $\mathrm{MSE}(\hat{\bar{y}}_{N_i}^{*\mathrm{EB}}) = \mathrm{E}\{(\hat{\bar{y}}_{N_i}^{*\mathrm{EB}} - \bar{y}_{N_i}^{*})^2\} = M_{1i}(\boldsymbol{\theta}) + M_{2i}(\boldsymbol{\theta}) + 2M_{3i}(\boldsymbol{\theta})$, where $M_{1i}(\boldsymbol{\theta}) = \mathrm{E}([N_i^{-1}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\hat{\boldsymbol{\theta}}\} - N_i^{-1}\sum_{j=1}^{N_i}y_{ij}^{*}]^2)$, and $M_{2i}(\boldsymbol{\theta}) = \mathrm{E}([N_i^{-1}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\hat{\boldsymbol{\theta}}\} - N_i^{-1}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\boldsymbol{\theta}\}]^2)$. and the cross term $M_{3i}(\boldsymbol{\theta}) = \mathrm{E}([N_i^{-1}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\hat{\boldsymbol{\theta}}\} - N_i^{-1}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\boldsymbol{\theta}\}][N_i^{-1}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\boldsymbol{\theta}\} - N_i^{-1}\sum_{j=1}^{N_i}y_{ij}^{*}])$. We consider the cross term $M_{3i}(\boldsymbol{\theta})$. Using double-conditioning,

$$\begin{aligned} M_{3i}(\boldsymbol{\theta}) &= \mathrm{E}\Bigg[\mathrm{E}\Big(\big[\frac{1}{N_i}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\hat{\boldsymbol{\theta}}\} - \frac{1}{N_i}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\boldsymbol{\theta}\}\big] \\ &\quad\times\big[\frac{1}{N_i}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\boldsymbol{\theta}\} - \frac{1}{N_i}\sum_{j=1}^{N_i}y_{ij}^{*}\big] \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\Big)\Bigg] \\ &= \mathrm{E}\Bigg[\Big(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\hat{\boldsymbol{\theta}}\} - \frac{1}{N_i}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\boldsymbol{\theta}\}\Big) \\ &\quad\times\mathrm{E}\Big(\big[\frac{1}{N_i}\sum_{j=1}^{N_i}\mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z});\boldsymbol{\theta}\} - \frac{1}{N_i}\sum_{j=1}^{N_i}y_{ij}^{*}\big] \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\Big)\Bigg] = 0. \end{aligned}$$

The leading term $M_{1i}(\boldsymbol{\theta})$ is the MSE of the MMSE predictor (4), which has the form $M_{1i}(\boldsymbol{\theta}) = \mathrm{E}[\mathrm{E}\{\bar{y}_{N_i}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\} - \bar{y}_{N_i}^{*}]^2 = \mathrm{E}[\mathrm{V}\{\bar{y}_{N_i}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\}]$, where

$$\mathrm{V}\{\bar{y}_{N_i}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\} = \mathrm{E}\left(\frac{1}{N_i^2}\sum_{j\in\bar{s}_i}\sum_{k\in\bar{s}_i}\big[\mathrm{E}\{y_{ij}^{*}y_{ik}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\} - \mathrm{E}\{y_{ij}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\}\mathrm{E}\{y_{ik}^{*} \mid (\boldsymbol{y}^{\boldsymbol{*}},\boldsymbol{z})\}\big]\right).$$

The form for $M_{1i}(\boldsymbol{\theta})$ in Theorem 2.2 now follows from the conditional distributions derived through the proof of Theorem 2.1. □

Before presenting the MSE estimator, we make a couple of remarks on Theorem 2.2. A proof that a cross-term analogous to $M_{3i}(\boldsymbol{\theta})$ for the area-level model is zero is presented on page 141-142 of Rao and Molina (2015). Jiang (2003) uses a decomposition similar to that in Theorem 2.2 for the MSE of an empirical Bayes small area predictor constructed under a unit-level generalized linear mixed model. Also, see Jiang and Lahiri (2006). The simulations in Section 3.2 provide numerical evidence that the MSE is dominated by the leading term $M_{1i}(\boldsymbol{\theta})$, although a formal proof of the orders of the terms in the MSE is beyond the scope of this manuscript.

The estimate of $M_{1i}(\boldsymbol{\theta})$ is an estimate of the observed conditional variance $V\{\bar{y}^*_{N_i} \mid (\boldsymbol{y^*}, \boldsymbol{z})\}$ obtained by substituting the unknown $\boldsymbol{\theta}$ with the MLE $\hat{\boldsymbol{\theta}}$ and using the Gauss-Hermite approximation (7) for expectations with respect to the distribution of $b_i \mid (\boldsymbol{y^*}, \boldsymbol{z})$. We call this estimator the "one-step" estimator of the leading term. The "one-step" estimator is defined as

$$\hat{M}_{1i}(\hat{\boldsymbol{\theta}}) = \frac{1}{N_i^2} \sum_{j \in \bar{s}_i} \sum_{k \in \bar{s}_i} \Big[\mathrm{E}_A\{y^*_{ij} y^*_{ik} \mid (\boldsymbol{y^*}, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\} - \mathrm{E}_A\{y^*_{ij} \mid (\boldsymbol{y^*}, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\} \mathrm{E}_A\{y^*_{ik} \mid (\boldsymbol{y^*}, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\} \Big], \quad (9)$$

where $\mathrm{E}_A\{y^*_{ij} y^*_{ik} \mid (\boldsymbol{y^*}, \boldsymbol{z}); \boldsymbol{\theta}\}$ approximates $\mathrm{E}\{y^*_{ij} y^*_{ik} \mid (\boldsymbol{y^*}, \boldsymbol{z}); \boldsymbol{\theta}\}$ by replacing $\mathrm{E}\{p_{ij}\eta(2b_i) \mid (\boldsymbol{y^*}, \boldsymbol{z}); \boldsymbol{\theta}\}$ with $\mathrm{E}_A\{p_{ij}\eta(2b_i) \mid (\boldsymbol{y^*}, \boldsymbol{z}); \boldsymbol{\theta}\}$ and $\mathrm{E}\{p_{ij}p_{ik}\eta(2b_i) \mid (\boldsymbol{y^*}, \boldsymbol{z}); \boldsymbol{\theta}\}$ with $\mathrm{E}_A\{p_{ij}p_{ik}\eta(2b_i) \mid (\boldsymbol{y^*}, \boldsymbol{z}); \boldsymbol{\theta}\}$ based on the Gauss-Hermite approximation (7).

Approximations to $M_{2i}(\boldsymbol{\theta})$ are difficult to derive due to the complexity of the derivatives involved. Therefore, use of the bootstrap, as defined in (10) below, to estimate $M_{2i}(\boldsymbol{\theta})$ is more practical. For $b = 1, \ldots, B$, we generate $\{y^{*(b)}_{ij} : i = 1, \ldots, D; j = 1, \ldots, N_i\}$ from the zero-inflated lognormal model (1) with the original parameter estimator $\hat{\boldsymbol{\theta}}$. We then obtain a bootstrap sample $\{y^{*(b)}_{ij} : i = 1, \ldots, D; j \in s_i\}$ where s_i is the index set of the originally observed sample units in area i. We let $\hat{\boldsymbol{\theta}}^{(b)}$ denote the vector of parameter estimators obtained from the b-th bootstrap sample. A bootstrap estimator of $M_{2i}(\boldsymbol{\theta})$ is defined by

$$\hat{M}^{\mathrm{Boot}}_{2i} = B^{-1} \sum_{b=1}^{B} (\hat{\bar{y}}^{*(b)\mathrm{EB}}_{N_i} - \hat{\bar{y}}^{*(b)\mathrm{MMSE}}_{N_i})^2, \quad (10)$$

where $\hat{\bar{y}}^{*(b)\mathrm{EB}}_{N_i}$ is the EB predictor of $\bar{y}^{*(b)}_{N_i}$ based on $\hat{\boldsymbol{\theta}}^{(b)}$ and $\hat{\bar{y}}^{*(b)\mathrm{MMSE}}_{N_i}$ is the minimum MSE predictor based on $\hat{\boldsymbol{\theta}}$. Both $\hat{\bar{y}}^{*(b)\mathrm{EB}}_{N_i}$ and $\hat{\bar{y}}^{*(b)\mathrm{MMSE}}_{N_i}$ are constructed with the b-th bootstrap sample. We define a semi-bootstrap estimator of the MSE of $\hat{\bar{y}}^{*\mathrm{EB}}_{N_i}$ by

$$\hat{M}^{\mathrm{Semi-Boot}}_{i} = \hat{M}_{1i}(\hat{\boldsymbol{\theta}}) + \hat{M}^{\mathrm{Boot}}_{2i}. \quad (11)$$

where $\hat{M}_{1i}(\hat{\boldsymbol{\theta}})$ is the proposed one-step MSE estimator as in (9). The MSE estimator (11) intentionally ignores the cross term $M_{3i}(\boldsymbol{\theta})$ in the decomposition of the MSE of $\hat{\bar{y}}^{*\mathrm{EB}}_{N_i}$ because this cross term is theoretically 0. Our simulation study confirms that the cross term is indeed negligible. Note that González-Manteiga et al. (2008) consider an estimator with a form analogous to the semi-boot estimator (11) in the context of a linear mixed effects model.

The semi-bootstrap estimator of the MSE does not incorporate an estimate of the bias of the one-step MSE estimator for the leading term in the MSE. This bias is defined formally as $\mathrm{E}\{\hat{M}_{1i}(\hat{\boldsymbol{\theta}}) - M_{1i}(\boldsymbol{\theta})\}$. Typically, in small area estimation, the double bootstrap is needed to estimate the bias (Hall and Maiti, 2006). Because we are able to calculate $M_{1i}(\boldsymbol{\theta})$ in one step, we can estimate the bias without requiring the double bootstrap. We assess the importance of this bias and compare the semi-boot estimator of the MSE to a fully parametric bootstrap MSE estimator in simulations.

3 Simulations

We evaluate the properties of the proposed EB predictor through Monte Carlo (MC) simulations.We simulate $M = 1000$ populations $\{y^*_{ij} : j = 1, ..., N_i; i = 1, ..., D\}$ from the zero-inflated lognormal model (1). We fix the parameters so that the proportion of zeros in the simulation is about 15%, similar to the CEAP data for South Dakota. A single set of covariates $\{z_{ij} : j = 1, ..., N_i; i = 1, ..., D\}$ is generated as mutually independent $N(4.45, 0.055)$ random variables and held fixed across simulations. The covariate z_{ij} is used for both model parts such that $\boldsymbol{z}_{1,ij} = \boldsymbol{z}_{2,ij} = z_{ij}$ The logit link is used for the binary part, and the parameters are $\boldsymbol{\beta} = (-13, 2)'$, $\boldsymbol{\alpha} = (-20, 5)'$, and $(\sigma^2_u, \sigma^2_e, \sigma^2_b) = (0.22, 1.23, 0.52)$. We vary the value of the correlation parameter ρ in the simulations below. There are 64 counties in the CEAP data analysis, so $D = 60$ areas are simulated. In each generated realization, $(N_i, n_i) = (71, 5)$ for 20 of the areas, $(143, 10)$ for another 20 of the areas and $(286, 20)$ for the remaining 20 areas so that $(N, n) = (10000, 700)$. In Section 3.1, we evaluate the effect of ignoring the correlation by comparing the EB predictor with simpler alternatives. In Section 3.2, we compare the proposed semi-boot MSE estimator to a more traditional bootstrap MSE estimator and assess the importance of the bias of the estimator of the leading term in the MSE.

3.1 Comparison with alternative predictors

The four alternative predictors to be compared with the EB predictor (8) are computationally simple but lack optimality. The alternative predictors use $\hat{\boldsymbol{\theta}}_0$, the estimator of $\boldsymbol{\theta}$ obtained by fitting the lognormal model (2) to the positive responses and independently fitting the model (3) to the $\{\delta_{ij} : i = 1, \ldots, D; j = 1, \ldots, n_i\}$. We use the `R` functions `lmer` and `glmer` in the package `lme4` (Bates et al., 2015) for the positive and binary parts, respectively. Restricted maximum likelihood (REML) is used to estimate the parameters of the lognormal component and maximum likelihood is used to estimate the parameters of the binary component. The first alternative predictor is the plug-in predictor (Chandra and Chambers, 2016) defined by replacing $\hat{y}^{*\mathrm{MMSE}}_{ij}(\hat{\boldsymbol{\theta}})$ in (8) with $\hat{y}^{*\mathrm{PI}}_{ij} = h_{ij}(\hat{\boldsymbol{\theta}}_0)\hat{p}^{\mathrm{PI}}_{ij}(\hat{\boldsymbol{\theta}}_0)$, where $\hat{p}^{\mathrm{PI}}_{ij}(\hat{\boldsymbol{\theta}}_0) = g^{-1}(\hat{\alpha}_0 + \boldsymbol{z}'_{2,ij}\hat{\boldsymbol{\alpha}}_1 + \hat{b}_i)$ and $\hat{b}_i$ is the predictor of b_i obtained from the penalized quasi-likelihood method of Schall (1991) implemented in the `R` function `glmer`. The second alternative, the "zero-ignored predictor", is the estimated MMSE predictor for the mean of the positive component defined in Berg and Chandra (2014) as $\hat{\bar{y}}^{*\mathrm{ZI}}_{N_i} = \tilde{N}_i^{-1}\{\sum_{j\in s_i} y^*_{ij} + \sum_{j\in \bar{s}_i} h_{ij}(\hat{\boldsymbol{\theta}}_0)\}$, where $\tilde{N}_i = N_i - \sum_{j\in s_i}(1 - \delta_{ij}) = N_i - (n_i - \tilde{n}_i)$ is the adjusted population size which excludes the number of zeros in the sample for area i. We refer to the predictor $\hat{\bar{y}}^{*\mathrm{ZI}}_{N_i}$ as the zero-ignored MMSE predictor. Rather than ignoring zeros completely, the third approach begins by adding a small positive constant ϵ, such as the minimum observed positive value, to y^*_{ij}. The third predictor, referred to as the shifted MMSE predictor, is defined by $\hat{\bar{y}}^{*\mathrm{SI}}_{N_i} = \max\{\hat{\bar{y}}^{*\mathrm{MMSE}}_{N_i,\epsilon} - \epsilon, 0\}$, where $\hat{\bar{y}}^{*\mathrm{MMSE}}_{N_i,\epsilon}$ is the MMSE predictor of Berg and Chandra (2014) applied to $y^*_{ij} + \epsilon$. Finally, we define the EB(0) predictor as the EB predictor constructed with $\hat{\boldsymbol{\theta}}_0$ and $\rho = 0$. Thus, EB(0) is an estimated MMSE predictor if the true correlation ρ is 0. The functional forms of the EB(0) and "plug-in" predictors differ only in the predictor of δ_{ij}.

We define the average MC MSE of the EB predictor by

$$\bar{\mathrm{MSE}}_{\mathrm{MC}}(\hat{\bar{y}}^{*\mathrm{EB}}_{N_i}) = \frac{1}{|\{k : N_k = N_i\}|}\sum_{\{k:N_k=N_i\}} M^{-1}\sum_{m=1}^{M}(\hat{\bar{y}}^{*(m)\mathrm{EB}}_{N_i} - \bar{y}^{*(m)}_{N_i})^2, \tag{12}$$

where $\hat{\bar{y}}^{*(m)\mathrm{EB}}_{N_i}$ is the EB predictor obtained in the m-th MC simulation, and $\bar{y}^{*(m)}_{N_i}$ is the corresponding population mean simulated in the m-th MC simulation. Note that the average in (12) is across MC iterations and areas with the common number of elements. We define the average MSE difference between an

alternative predictor (EB(0), PI, ZI and SI) and the EB predictor as

$$\text{MSEDiff}_{\text{MC}}(\hat{\bar{y}}_{N_i}^{*\text{Alt}}) = \frac{1}{|\{k : N_k = N_i\}|} \sum_{\{k:N_k=N_i\}} M^{-1} \sum_{m=1}^{M} (\hat{\bar{y}}_{N_i}^{*(m)\text{Alt}} - \bar{y}_{N_i}^{*(m)})^2 - \bar{\text{MSE}}_{\text{MC}}(\hat{\bar{y}}_{N_i}^{*\text{EB}}), \tag{13}$$

where $\hat{\bar{y}}_{N_i}^{*(m)\text{Alt}}$ is an alternative predictor of the i-th area mean in the m-th MC simulation. To account for the variance of $\text{MSEDiff}_{\text{MC}}(\hat{\bar{y}}_{N_i}^{*\text{Alt}})$ as a MC approximation for the true difference, we define a margin of error as 1.96 times the MC standard error. (Detailed definition in Section S3 in Supporting Information.) A 95% confidence interval for each average MSE difference $\bar{\zeta}$ is given by $(\bar{\zeta} - \omega, \bar{\zeta} + \omega)$ where ω denotes the corresponding margin of error.

Table 1 reports the average MSE difference (multiplied by 10^5) between each alternate predictor and the EB predictor for $\rho = 0.9$. As a reference, the average MC MSEs (multiplied by 10^5) of the EB predictor defined in (12) are 24.09, 15.92 and 9.27, respectively for $n_i = 5, 10$ and 20. The relative MSE difference (ratio of the average MSE difference to the average MSE of the EB predictor) tends to increase as the area sample size increases. The zero-ignored MMSE predictor is clearly inefficient with relative MSE differences between 18% and 30%, implying that simply tossing the zeros is not a good approach. Interestingly, the shifted MMSE predictor introduces an even larger error by adding a small positive constant before the log-transformation and corrupting the normality. As a consequence of mis-specification of the correlation structure, the relative MSE differences for the EB(0) and "plug-in" predictors are about 5% to 7%. The margins of error in Table 1 indicate that the increase in MSE of the EB(0) and "plug-in" predictors relative to the EB predictor is significant.

Given that the EB(0) predictor is most competitive with the EB predictor when $\rho = 0.9$, we next assess the effect of the size of ρ on the relative efficiency of the EB predictor to the EB(0) predictor. Table 2 contains the average MSE differences (multiplied by 10^5) between the EB(0) and EB predictor for values of ρ ranging from -0.9 to 0.9. When $|\rho| = 0.3$, the MSE of the EB(0) predictor does not differ significantly from that of the EB predictor. As $|\rho|$ increases to 0.6 or 0.9, the EB(0) predictor has larger MSE than the EB predictor. Because the EB predictor is over-parametrized when $\rho = 0$, the EB(0) predictor is superior to the EB predictor when the true correlation is indeed zero.

The comparison to alternative predictors above focuses on predictors constructed under the assumptions of a lognormal model. The remaining predictor for zero-inflated skewed data in the small area literature to which we do not compare is the zero-inflated gamma model presented in Dreassi et al. (2014). We do not include their model in our study because the lognormal and gamma models are two different models for the response distribution. We expect predictors constructed under the assumptions of the lognormal model to outperform predictors constructed under the assumption of a gamma model if the lognormal model is true (and vice versa). We consider an investigation into properties of predictors constructed under a misspecified model assumption to be an area for future work. We focus on predictors constructed under the assumptions of the lognormal model because the lognormal model appears adequate for the CEAP data, as discussed further in Section 4.

3.2 Evaluation of the parametric bootstrap MSE terms

We use a traditional parametric bootstrap method to define another MSE estimator besides the one-step and semi-boot MSE estimators. A fully parametric bootstrap MSE estimator is defined by

$$\hat{M}_i^{\text{Boot}} = B^{-1} \sum_{b=1}^{B} (\hat{\bar{y}}_{N_i}^{*(b)\text{EB}} - \bar{y}_{N_i}^{*(b)})^2, \tag{14}$$

where $\hat{\bar{y}}_{N_i}^{*(b)\text{EB}}$ is the same as defined in (10) and $\bar{y}_{N_i}^{*(b)} = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}^{*(b)}$ is the i-th small area mean based on the b-th bootstrap population. Fully parametric bootstrap estimators such as (14) are suggested

in Molina and Rao (2010) for transformed parameters, in González-Manteiga et al. (2007) for a logistic mixed effects model, and in González-Manteiga et al. (2008) for a linear mixed effects model.

The bootstrap MSE estimator (14) of the EB predictor can be decomposed into three parts as $\hat{M}_i^{\text{Boot}} = \hat{M}_{1i}^{\text{Boot}} + \hat{M}_{2i}^{\text{Boot}} + 2\hat{M}_{3i}^{\text{Boot}}$, where $\hat{M}_{1i}^{\text{Boot}} = B^{-1}\sum_{b=1}^{B}(\hat{\bar{y}}_{N_i}^{*(b)\text{MMSE}} - \bar{y}_{N_i}^{*(b)})^2$, $\hat{M}_{2i}^{\text{Boot}}$ is defined as in (10) and a bootstrap estimator of the cross term $\hat{M}_{3i}^{\text{Boot}} = B^{-1}\sum_{b=1}^{B}(\hat{\bar{y}}_{N_i}^{*(b)\text{MMSE}} - \bar{y}_{N_i}^{*(b)})(\hat{\bar{y}}_{N_i}^{*(b)\text{EB}} - \hat{\bar{y}}_{N_i}^{*(b)\text{MMSE}})$. A bootstrap estimator of the bias of the estimator of the leading term in the MSE is defined by $\hat{M}_{1i}^{\text{Bias}} = B^{-1}\sum_{b=1}^{B}\hat{M}_{1i}(\hat{\boldsymbol{\theta}}^{(b)}) - \hat{M}_{1i}(\hat{\boldsymbol{\theta}})$, where $\hat{M}_{1i}(\hat{\boldsymbol{\theta}}^{(b)})$ is the one-step MSE estimator constructed with the original sample and the b-th bootstrap parameter estimate $\hat{\boldsymbol{\theta}}^{(b)}$. The semi-boot MSE estimator (11) intentionally ignores the cross-term $M_{3i}(\boldsymbol{\theta})$ because the expected value of the cross-term is zero, as explained in the proof of Theorem 2.2. The one-step MSE estimator (9) additionally ignores M_{2i}, the error due to estimating model parameters. Both the fully parametric and semi-boot MSE estimator ignore M_{1i}^{Bias}, the bias of the estimator of the leading term. Therefore, an assessment of the importance of the three MSE components M_{3i}^{Boot}, M_{2i} and M_{1i}^{Bias} is relevant.

The MSE components and MSE estimators are evaluated through simulations, where the simulation configuration is the same as Section 3.1 with $\rho = 0.9$, a MC sample size of 1000, and a bootstrap sample size of $B = 100$ in each MC sample. We group by sample size and compute the MC means of the ratios of $\hat{M}_{3i}$, $\hat{M}_{2i}$, and $\hat{M}_{1i}^{\text{Bias}}$ to the fully parametric bootstrap MSE estimator (14). (We defer full tabular output to Table S2 in Supporting Information.) Accounting for parameter estimation through $\hat{M}_{2i}$ accounts for about 5-6% of the total MSE. The bias of estimating the leading term is less important, as $\hat{M}_{1i}^{\text{Bias}}$ accounts for -2.14%, -2.98%, and -3.06% of $\hat{M}_i^{\text{Boot}}$ on average when the sample size is 5, 10, and 20, respectively. The average ratios for the cross-term $\hat{M}_{3i}$ are approximately -0.2%, supporting the theory that the cross-term is negligible.

We next consider the biases for the MC MSE (13) and empirical coverages of normal theory prediction intervals for the three MSE estimators (one-step, bootstrap and semi-boot). The analysis of the MC bias and empirical coverage presented in Table 3 provides further support for the semi-boot MSE estimator. Accounting for parameter estimation significantly reduces the negative bias of the one-step MSE estimator and keeps the coverage rate close to the nominal 95% level. The bootstrap MSE estimator is unbiased but has significantly lower coverage than the one-step MSE estimator. We believe that the semi-boot MSE estimator is superior to the fully parametric bootstrap MSE estimator because the fully parametric bootstrap MSE estimator unnecessarily estimates the cross-term.

4 Estimating sheet and rill erosion for the Conservation Effects Assessment Project

The Conservation Effects Assessment Project (CEAP), detailed in Goebel (2012), is a collection of four types of surveys that assess conditions on cropland, grazing lands, wildlife, and wetlands. The cropland assessment, the focus of our work, monitors soil and chemical loss from crop fields. The sample for the cropland assessment is a subset of locations classified as cropland in a larger survey called the National Resources Inventory (Nusser and Goebel, 1997). Policies that allocate resources at a local level can benefit from small area estimates of variables collected in CEAP. Therefore, we consider county estimation for CEAP in South Dakota.

The response variable of interest for our study (y^*) is a measure of sheet and rill erosion in tons per year obtained from the Revised Universal Soil Loss Equation, version 2 (RUSLE2), one of the outputs of the Agricultural Policy/Environmental eXtender (APEX) model (Williams and Izaurralde, 2006). RUSLE2 is intended to improve upon the Universal Soil Loss Equation (USLE), a traditional measure of sheet and rill erosion, published in Wischmeir and Smith (1965). USLE includes several fundamental factors accounting for soil erosion, some of which we use in our model for the computationally more complex

RUSLE2. Specifically, the USLE equation is defined by

$$\mathrm{A} = \mathrm{R\ K\ L\ S\ C\ P}, \tag{15}$$

where A is USLE soil loss per unit area, R is the rainfall factor, K is the soil erodibility index, L is the slope-length, S is the slope-gradient, C is the cropping-management effect relative to a fallow field, and P is the erosion-control practice factor. The USLE equation (15) implies a log-linear relationship between the sheet and rill erosion and the USLE factors. RUSLE2 maintains many of the fundamental concepts of the USLE equation but uses more granular weather information and a more detailed cover-management subfactor.

The proportions of non-zeros and sample sizes of the CEAP RUSLE2 data collected in the 66 counties of South Dakota tend to increase from west to east, as shown in Figure S2 of the Supporting Information. The east of South Dakota has more sampled units than the west, because most of the cropland in South Dakota is located to the east of the Missouri river. The sample size is less than 5 for most of the counties in the west. Butte county and Lawrence county are considered out of scope because 80% of the lands in those two counties are rangeland or federal land. Among the 64 counties in the cropland population of interest, seven of them have no positive RUSLE2 measurement. The overall proportion of zeros in the sampled cropland RUSLE2 measurements is 15%.

4.1 Auxiliary variables and population predictions for CEAP

We require auxiliary variables that are known for the full population of cropland locations in South Dakota. A simple and complete list of crop locations in South Dakota does not exist. Compiling the necessary auxiliary information requires combining three additional sources besides CEAP: National Resources Inventory (NRI), Soil Survey Geographic Database (SSURGO) and Cropland Data Layer (CDL). We use these sources for two purposes. The first is modeling cropland RUSLE2. The second is defining the full population of crop fields for which predictions are required. These three sources provide covariates that are known for the full population of cropland in South Dakota and are related to the USLE factors.

The first covariate is the log of the USLE R-factor, $\log(R)$. We obtain the USLE rainfall factor from NRI. The R-factor is nearly constant within a county due to the uniform climate conditions in the midwestern United States. We define a county level auxiliary variable by the the most frequently recorded NRI R-factor in each county of South Dakota.

We obtain covariates for USLE K and S factors from SSURGO (U.S. Department of Agriculture, 2020*b*), a detailed map with soil properties for mapunits covering most of the United States. One of the properties is KWFACT, an erosion factor modified by the presence of the rock fragment and ranging from 0.02 to 0.69. We use the values of $\log(K)$, the log of the KWFACT as approximations for the log of the USLE K-factors. Another property is SLOPE_R, the representative slope gradient in percentage. To convert slope gradient to a USLE S-factor, Smith and Wischmeier (1957) proposed $S = (0.43 + 0.30s + 0.043s^2)/6.613$, where s is the slope gradient and S is the USLE S-factor. We use $\log(S)$ as a covariate.

Cropland Data Layer (U.S. Department of Agriculture, 2020*a*) is a geo-referenced and crop-specific land cover data layer. Given the CEAP sample is collected by the 2003-2006 CEAP survey and South Dakota was not included in the CDL project until 2006, we use the 2006 CDL data for South Dakota. Corn, soybean, spring wheat and winter wheat are the four dominant crops in South Dakota according to the 2006 CDL. Thus we classify each 2006 CDL pixel at a spatial resolution of 56 meters into one of these four categories or the "remainder" category. The crop-based CDL classification is our best available approximation to the USLE C-factor.

These three sources (NRI, SSURGO and CDL) result in seven possible explanatory variables: $\log(R)$, $\log(K)$, $\log(S)$, and four indicator variables for CDL membership in corn, soybeans, spring wheat, or winter wheat. (The baseline category is the remainder of CDL crop categories.) These variables provide $\boldsymbol{z}_{ij}$ in the small area model. Although possible covariates are chosen to be related to USLE factors,

RUSLE2 is a complicated function of a lot of factors which may not be available for the full population. The factors we represent through possible covariates are those we are able to populate at each cropland location.

The population of interest is the collection of all crop fields in South Dakota. We define the population of interest to be locations with CDL pixels that are classified as cropland according to the NRI definition of cropland (NRI Broad-use Category 1 or 2, defined in terms of specific crops in Table S3 of the Supporting Information). Our constructed population consists of about 20 million CDL pixels of cropland in 6,615 soil map units. Several soil map units overlap with more than one county, so we overlaid those map unit polygons with the county shapefiles to obtain their intersections. After excluding one sampled point which falls out of the 6,615 soil map units classified as eligible, the sample dataset contains a total of $n = 641$ different geographic locations in the CEAP sample for South Dakota.

Obtaining the soil map unit classification and subsequently defining a prediction for 20 million distinct CDL pixels is computationally prohibitive. To facilitate the computations, we exploit the fact that the predictors are constant within an intersection of a soil map unit, CDL crop type, and county. This simplification allows us to express the predictor as a weighted sum of a smaller number of distinct entities. We refer the reader to Section S1 of the Supporting Information for an explanation of how we incorporate weights to reduce the computational complexity for the CEAP analysis.

4.2 CEAP zero-inflated lognormal model fitting

We fit the zero-inflated lognormal model (1) with the logit link to the observed data $(\boldsymbol{y}^*, \boldsymbol{z})$. To select the covariates, we apply backward variable selection to the model (2) for the positive part and to the model (3) for the binary part separately. We use $\Delta(\text{AIC}) = 0.5$ as a threshold for variable exclusion. As shown in Table 4, the best predictors according to this criterion are *logR*, *logK* together with *logS* for the positive part and *logR*, *logS*, *is.soybean* together with *is.sprwht* for the binary part. When the zero-inflated lognormal model assuming correlation between the random effects with all possible explanatory variables is fit to the CEAP RUSLE2 data, the coefficients associated with the selected covariates remain significant. Table 4 contains the maximum likelihood estimates of the coefficients and the corresponding bootstrap standard errors based on the assumed model with estimated ρ. A nominal 95% confidence interval for ρ based on the lower and upper 2.5% percentiles of parametric bootstrap distribution of $\hat{\boldsymbol{\theta}}$ (with bootstrap sample size of 500) is $(0.21, 0.99)$. In consistency with the soil loss equation (15), the R-factor, K-factor and S-factor are positively related with soil loss per year. The probability that cropland RUSLE2 is positive, or there exists soil loss, is estimated to be significantly higher for the crop type of soybean or spring wheat relative to the remainder category.

The simple logit link and log transformations used to fit the model are special cases of parametric families of link functions and transformations. We considered alternate link functions besides the logit for the binary part and transformations in the family of Box-Cox power transformations (Box and Cox, 1964) for the positive part. The analysis detailed in Section S2 of the Supporting Information provides no evidence against the simple choices of the logit link and logarithmic transformation. We therefore use the logit link and the logarithmic transformation for the CEAP data analysis.

We assess the goodness of fit of the model using residuals for the positive part and a Hosmer-Lemeshow test for the binary part. We constructed marginal and conditional residuals defined as

$$r_{ij,\text{marg}} = (\log(y_{ij}) - \boldsymbol{z}'_{1,ij}\hat{\boldsymbol{\beta}})/\sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}$$

and as

$$r_{ij,\text{cond}} = (\log(y_{ij}) - \boldsymbol{z}'_{1,ij}\hat{\boldsymbol{\beta}} - \hat{E}[u_i \mid (\boldsymbol{y}^*, \boldsymbol{z})])/\sqrt{\hat{\sigma}_e^2}$$

respectively, where $\hat{\boldsymbol{\beta}}$ is the MLE estimate of the regression coefficient for the positive part and $\hat{E}[u_i \mid (\boldsymbol{y}^*, \boldsymbol{z})]$ denotes the estimated conditional expected value of u_i given the data (after integrating over the conditional distribution of b_i). We plotted the residuals against $\boldsymbol{z}'_{1,ij}\hat{\boldsymbol{\beta}}$ and against $\boldsymbol{z}'_{1,ij}\hat{\boldsymbol{\beta}} + \hat{E}[u_i \mid (\boldsymbol{y}^*, \boldsymbol{z})]$.

The residual plots (presented in Figure S3 of the Supporting Information) reveal that the RUSLE2 values are rounded so that the same values of $\log(y_{ij})$ are duplicated in the data set. Otherwise, the residual plots show no important trends as a function of the fitted values. The p-values of Shaprio-Wilk tests for normality exceed 0.05, providing essentially no evidence to reject the normality assumption. For the binary part, we construct Hosmer-Lemeshow tests (Hosmer Jr and Lemeshow, 2000) using an estimate of $E[p_{ij} \mid (\boldsymbol{y}^*, \boldsymbol{z})]$ as the predicted probability. The p-values (presented specifically in Figure S4 of the Supporting Information) for group sizes ranging from 5 to 15 are all above 0.05. These analyses provide support for the zero-inflated lognormal model.

4.3 CEAP empirical Bayesian predictions

We obtain the EB predicted mean cropland RUSLE2 in the 64 sampled counties of South Dakota with the fitted zero-inflated lognormal model with correlated random effects. We compare the EB predictors and corresponding MSE estimators to direct estimators and associated standard errors. The direct estimator is the sample mean. Scatterplots of the EB predictors (presented in Figure S5 of the Supporting Information) against the direct estimators show a strong, linear association. The strength of the association increases as the area sample size increases, an expected trend because increasing the sample size should reduce the effect of modeling on the predictors. We define a standard error of the direct estimator by $\mathrm{SE}_{\mathrm{pool}}(\hat{\bar{y}}^*_{N_i}) = \sqrt{S^2/n_i}$, where S^2 is the pooled within-county variance given by $S^2 = (n-D)^{-1}\sum_{i=1}^{D}\sum_{j=1}^{n_i}(y^*_{ij} - \bar{y}^*_{i\cdot})^2$ and $\bar{y}^*_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} y^*_{ij}$. We use the pooled variance S^2 because the sample variance for the i-th county is undefined if the sample size n_i is 1. Figure 2 shows the standard errors of the direct estimators along with the square root of the one-step MSE estimator and the square root of the semi-boot MSE estimator constructed with bootstrap size $B = 100$. The standard errors computed by the one-step method are close to the semi-boot method. With either the one-step or the semi-boot standard error, as depicted in Figure 2, the proposed EB predictor is more precise than the direct estimator across different sample sizes. The reduction in standard error tends to be greater for the group of counties with smaller sample sizes.

Figure 3 gives a cartogram of the EB predicted mean cropland RUSLE2 in the 64 sampled counties of South Dakota. The darkness of the shade corresponds to the magnitude of the prediction. The relative size of the shaded region to the total area of the county is inversely related to the estimated coefficient of variation (CV), a graphical technique intended to draw the reader's attention toward more reliable estimates. The predicted erosion tends to be highest in magnitude and most reliable (smallest estimated CV) in the eastern, particularly the southeastern, portion of South Dakota. The increase in predicted erosion from west to east occurs partly because rainfall and the prevalence of soybeans are greater in the east than in the west of South Dakota. The decrease in estimated CV from west to east likely occurs because the CEAP sample is more dense in more highly cultivated areas.

5 Conclusion

A zero-inflated lognormal model with correlated random area effects has been proposed for modeling soil loss data, and empirical Bayesian small area estimators have been derived from the model. By deriving the conditional distribution of the random components in the model, we are able to use the maximum likelihood method to estimate the whole parameter vector even with the correlation coefficient ρ. The model assumptions appear reasonable for the data analysis of the cropland CEAP RUSLE2 measurements collected in South Dakota. The correlation of the random area effects between the two parts is estimated to be positive and moderately large. The model based EB predictors of area means have values near the direct estimator (sample mean) when the sample size is relatively large. The standard errors of the EB predictor are estimated to be much smaller than the direct estimator, especially for counties with only a few samples.

In the simulation study, we compare our proposed EB predictor of area means with the EB(0), plug-in, zero-ignored MMSE and shifted MMSE predictor. Simply tossing all the zeros in the sample would

result in poor estimation results, and shifting the responses by the amount of the smallest positive quantity leads to even larger mean square errors. The plug-in estimator has comparable performance with the EB(0) predictor, both of which assume independence between the two parts and have moderately larger MSE than the proposed EB predictor when the independence assumption is violated. For the proposed EB predictor, we give three ways to estimate the mean square error: an analytic estimator and two bootstrap estimators. The analytic approximation is based on the observed conditional variance. The two bootstrap MSE estimators take into account the variance due to parameter estimation. The semi-bootstrap MSE estimator that adds the variance due to parameter estimation to the estimator of the conditional variance outperforms the other MSE estimators. This MSE estimator differs from the full parametric bootstrap MSE estimator because it uses the fact that the prediction error is uncorrelated with the difference between the EB predictor and the MMSE predictor. For both the simulation and the data analysis, the sample size is large enough that the observed conditional variance and the semi-boot MSE estimator are close.

In the supporting information, we also show that our proposed estimators can accommodate any parametrized link functions for the binary part of the model as well as parametrized transformations for the positive part. Although a generalized link function for the positive part may prevent one from using the analytic form of the proposed EB predictor in the paper, the conditional distributions of the random components are implicitly using the link functions. Therefore, one can easily adapt the simulation-based method of Molina and Rao (2010) to obtain the predictions. In the data analysis, it is shown that the proposed predictor can be easily adapted to incorporate weights for averaging across the individual level predicted values.

The data analysis also suggests areas for future work. The observed cropland RUSLE2 measurements are rounded to the three decimal places. Modifying our EB approach to account for the discrete nature of the data is an area for future study. The lognormal model for the positive component, though reasonable for South Dakota, may not hold for all states and response variables of interest. Investigation of compromises between the fully parametric approach considered here and the semiparametric quantile regression model of Berg and Lee (2019) is a possible avenue for future research.

Acknowledgements The research was partially supported by a grant from US NSF (MMS-1733572) and a cooperative agreement between Iowa State University and USDA-NRCS.

Conflict of Interest

The authors have declared no conflict of interest.

Appendix

To maximize the likelihood, we use the method `BFGS` in the `R` function `optim`, which is a quasi-Newton algorithm that uses information about the objective function and its gradient. To improve the speed of the optimization, we supply `optim` with a function to calculate the gradient vector to find the fastest ascending direction. The functional form for the gradient vector is given in subsection A.1 below. In the `R` code, we express the functions defining the log likelihood and its gradient in terms of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_e, \sigma_u, \sigma_b, t)'$ where $t = \tan(\rho\pi/2)$. This transformation for ρ avoids a need to specify bounds for the parameter space. To obtain starting values for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, σ_e, σ_u, and σ_b, we begin with the elements of $\hat{\boldsymbol{\theta}}_0$, the parameter estimator from fitting the two model parts independently as described in Section 3.1. We then bound the starting value of σ_b below by $\sqrt{10^{-5}}$ because the gradient vector is undefined if the starting value for σ_b is zero. The `R` functions `lmer` and `glmer` furnish predictors of u_i and of b_i, where the predictor of u_i is the best linear unbiased predictor, and the predictor of b_i is calculated using PQL. We use as the starting value for ρ the sample correlation between the initial predictors of b_i and u_i. If the starting value for σ_b is set to the bound of $\sqrt{10^{-5}}$, the starting value for ρ is set to zero.

A.1 Gradient vector

We give the functional form for the gradient of the log likelihood $l_i(\boldsymbol{\theta}) = \log(L_i(\boldsymbol{\theta}))$. Denote $\boldsymbol{Z}_{1i}$ as the model matrix and $\tilde{\boldsymbol{r}}_i = (\tilde{r}_{i1}, \ldots, \tilde{r}_{in_i})'$ as marginal residual vector of model (2) for area i. It can be shown that $\partial\bar{\tilde{r}}_i/\partial\boldsymbol{\beta} = -\bar{\tilde{\boldsymbol{z}}}_{1i}$, $\partial m_i/\partial\boldsymbol{\beta} = m_i/\bar{\tilde{r}}_i\bar{\tilde{\boldsymbol{z}}}_{1i}$ where $\bar{\tilde{\boldsymbol{z}}}_{1i} = \tilde{n}_i^{-1}\boldsymbol{Z}_{1i}'\boldsymbol{\delta}_i$ and $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{in_i})'$. Denote $\boldsymbol{Z}_{2i}$ as the model matrix of model (3) for area i. It can be shown that $\partial\pi_{s_i}/\partial\boldsymbol{\alpha} = \pi_{s_i}(b_i)\boldsymbol{Z}_{2i}'\boldsymbol{u}_i$, where $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{in_i})'$, $u_{ij} = (g^{-1})'(\alpha_0 + \boldsymbol{z}_{2,ij}'\boldsymbol{\alpha}_1 + b_i)p_{ij}^{-\delta_{ij}}(1-p_{ij})^{\delta_{ij}-1}(2\delta_{ij}-1)$ and $(g^{-1})'(\cdot)$ is the derivative function of $g^{-1}(\cdot)$. When $g(\cdot)$ is the logit function, u_{ij} simplifies to $p_{ij}^{1-\delta_{ij}}(1-p_{ij})^{\delta_{ij}}(2\delta_{ij}-1)$. The derivatives of the log likelihood function with respect to the fixed effect coffcient vectors are

$$\begin{aligned}
\frac{\partial l_i}{\partial\boldsymbol{\beta}} &= -\frac{\gamma_i\bar{\tilde{r}}_i}{\sigma_e^2/\tilde{n}_i}\bar{\tilde{\boldsymbol{z}}}_{1i} - \frac{m_i}{v_i}\frac{\partial m_i}{\partial\boldsymbol{\beta}} + \frac{1}{\sigma_e^2}\boldsymbol{Z}_{1i}'\tilde{\boldsymbol{r}}_i + \frac{\int \pi_{s_i}(b_i)(-\frac{b_i-m_i}{\sqrt{v_i}})\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}\frac{1}{\sqrt{v_i}}\frac{\partial m_i}{\partial\boldsymbol{\beta}} \\
\frac{\partial l_i}{\partial\boldsymbol{\alpha}} &= \frac{\int(\frac{\partial\pi_{s_i}}{\partial\boldsymbol{\alpha}})\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}.
\end{aligned}$$

For derivatives with respect to the variance components, it can be derived that $\partial\gamma_i/\partial\sigma_u = 2\gamma_i(1-\gamma_i)/\sigma_u$, $\partial\gamma_i/\partial\sigma_e = -2\gamma_i(1-\gamma_i)/\sigma_e$, $\partial m_i/\partial\sigma_u = (m_i/\sigma_u)(\sigma_e^2/\tilde{n}_i - \sigma_u^2)/(\sigma_e^2/\tilde{n}_i + \sigma_u^2)$, $\partial m_i/\partial\sigma_b = m_i/\sigma_b$, $\partial m_i/\partial\sigma_e = -2m_i(\sigma_e/\tilde{n}_i)/(\sigma_e^2/\tilde{n}_i + \sigma_u^2)$, $\partial v_i/\partial\gamma_i = -v_i\rho^2/\{1-(1-\gamma_i)\rho^2\}$, $\partial v_i/\partial\sigma_b = 2v_i/\sigma_b$. The derivatives of v_i use the fact that $v_i = \sigma_b^2(1-\rho^2)/\{1-(1-\gamma_i)\rho^2\}$ is an equivalent expression of v_i as in Theorem 2.1. Then

$$\begin{aligned}
\frac{\partial l_i}{\partial\sigma_u} &= \frac{1}{2}\Big[\frac{-1}{1-\gamma_i}\frac{\partial\gamma_i}{\partial\sigma_u} + \frac{\bar{\tilde{r}}_i^2}{\sigma_e^2/\tilde{n}_i}\frac{\partial\gamma_i}{\partial\sigma_u} + \frac{2m_i}{v_i}\frac{\partial m_i}{\partial\sigma_u} - \frac{m_i^2}{v_i^2}\frac{\partial v_i}{\partial\sigma_u}\Big] \\
&\quad + \frac{\int \pi_{s_i}(b_i)[-(\frac{b_i-m_i}{\sqrt{v_i}})\frac{\partial}{\partial\sigma_u}(\frac{b_i-m_i}{\sqrt{v_i}})]\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i} \\
\frac{\partial l_i}{\partial\sigma_e} &= \frac{1}{2}\Big[\frac{-1}{1-\gamma_i}\frac{\partial\gamma_i}{\partial\sigma_e} - \frac{2}{\sigma_e/\tilde{n}_i} + \frac{\bar{\tilde{r}}_i^2}{\sigma_e^2/\tilde{n}_i}\frac{\partial\gamma_i}{\partial\sigma_e} - \frac{2\gamma_i\bar{\tilde{r}}_i^2}{\sigma_e^3/\tilde{n}_i} \\
&\quad + \frac{2m_i}{v_i}\frac{\partial m_i}{\partial\sigma_e} - \frac{m_i^2}{v_i^2}\frac{\partial v_i}{\partial\sigma_e} + 2\sigma_e^{-3}\sum_j \tilde{r}_{ij}^2\Big] \\
&\quad + \frac{\int \pi_{s_i}(b_i)[-(\frac{b_i-m_i}{\sqrt{v_i}})\frac{\partial}{\partial\sigma_e}(\frac{b_i-m_i}{\sqrt{v_i}})]\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i} \\
\frac{\partial l_i}{\partial\sigma_b} &= \frac{1}{2}\Big[\frac{2m_i}{v_i}\frac{\partial m_i}{\partial\sigma_b} - \frac{m_i^2}{v_i^2}\frac{\partial v_i}{\partial\sigma_b} - \frac{2}{\sigma_b}\Big] \\
&\quad + \frac{\int \pi_{s_i}(b_i)[-(\frac{b_i-m_i}{\sqrt{v_i}})\frac{\partial}{\partial\sigma_b}(\frac{b_i-m_i}{\sqrt{v_i}})]\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}
\end{aligned}$$

where $\partial\{(b_i-m_i)/\sqrt{v_i}\}/\partial\sigma = -\{\partial m_i/\partial\sigma + (b_i-m_i)/(2v_i)\partial v_i/\partial\sigma\}/\sqrt{v_i}$ for $\sigma = \sigma_u, \sigma_e$ or σ_b. For derivative with respect to ρ, we have $\partial\gamma_i/\partial\rho = -2\gamma_i(1-\gamma_i)\rho/(1-\rho^2)$, $\partial v_i/\partial\rho = -2\rho\gamma_i\sigma_b^2/\{1-(1-\gamma_i)\rho^2\}$, $\partial m_i/\partial\rho = \bar{\tilde{r}}_i\sigma_u\sigma_b/(\sigma_u^2 + \sigma_e^2/\tilde{n}_i)$. Then

$$\begin{aligned}
\frac{\partial l_i}{\partial\rho} &= \frac{1}{2}\Big[\frac{-1}{1-\gamma_i}\frac{\partial\gamma_i}{\partial\rho} + \frac{\bar{\tilde{r}}_i^2}{\sigma_e^2/\tilde{n}_i}\frac{\partial\gamma_i}{\partial\rho} + \frac{2m_i}{v_i}\frac{\partial m_i}{\partial\rho} - \frac{m_i^2}{v_i^2}\frac{\partial v_i}{\partial\rho}\Big] \\
&\quad + \frac{\int \pi_{s_i}(b_i)[-(\frac{b_i-m_i}{\sqrt{v_i}})\frac{\partial}{\partial\rho}(\frac{b_i-m_i}{\sqrt{v_i}})]\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}
\end{aligned}$$

where $\partial\{(b_i - m_i)/\sqrt{v_i}\}/\partial\rho = -\{\partial m_i/\partial\rho + (b_i - m_i)/(2v_i)\partial v_i/\partial\rho\}/\sqrt{v_i}$. The implementation of this MLE method inputs $t = \tan(\rho\pi/2)$ instead of ρ to the `optim` function. Therefore, the last element of the gradient vector is $\partial l_i/\partial t = (\partial l_i/\partial\rho)(\partial\rho/\partial t)$ where $\partial\rho/\partial t = 2/\{\pi(1+t^2)\}$.

References

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015), 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software* **67**(1), 1–48.

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**(401), 28–36.

Berg, E. and Chandra, H. (2014), 'Small area prediction for a unit-level lognormal model', *Computational Statistics & Data Analysis* **78**, 159–175.

Berg, E., Chandra, H. and Chambers, R. (2016), *Small Area Estimation for Lognormal Data*, John Wiley Sons, Ltd, chapter 15, pp. 279–298.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118814963.ch15*

Berg, E. and Lee, D. (2019), 'Small area prediction of quantiles for zero-inflated data and an informative sample design', *Statistical Theory and Related Fields* **3**(2), 114–128.

Box, G. E. and Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society: Series B (Methodological)* **26**(2), 211–243.

Chandra, H. and Chambers, R. (2016), 'Small area estimation for semicontinuous data', *Biometrical Journal* **58**(2), 303–319.

Dreassi, E., Petrucci, A. and Rocco, E. (2014), 'Small area estimation for semicontinuous skewed spatial data: An application to the grape wine production in tuscany', *Biometrical Journal* **56**(1), 141–156.

Erciulescu, A. L. and Fuller, W. A. (2016), 'Small area prediction under alternative model specifications', *Statistics in Transition new series* **17**(1), 9–24.

Fay III, R. E. and Herriot, R. A. (1979), 'Estimates of income for small places: an application of james-stein procedures to census data', *Journal of the American Statistical Association* **74**(366a), 269–277.

Giner, G. and Smyth, G. K. (2016), 'statmod: probability calculations for the inverse gaussian distribution', *R Journal* **8**(1), 339–351.

Goebel, J. (2012), 'Statistical methodology for the NRI-CEAP cropland survey'.
URL: *https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/16/nrcs143_013402.pdf*

Golub, G. H. and Welsch, J. H. (1969), 'Calculation of gauss quadrature rules', *Mathematics of computation* **23**(106), 221–230.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2007), 'Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model', *Computational statistics & data analysis* **51**(5), 2720–2733.

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

Hall, P. and Maiti, T. (2006), 'On parametric bootstrap methods for small area prediction', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(2), 221–238.

Hobza, T. and Morales, D. (2016), 'Empirical best prediction under unit-level logit mixed models', *Journal of official statistics* **32**(3), 661–692.

Hobza, T., Morales, D. and Santamaría, L. (2018), 'Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models', *Test* **27**(2), 270–294.

Hosmer Jr, D. W. and Lemeshow, S. (2000), *Applied logistic regression*, John Wiley & Sons.

Jiang, J. (2003), 'Empirical best prediction for small-area inference based on generalized linear mixed models', *Journal of Statistical Planning and Inference* **111**(1-2), 117–127.

Jiang, J. and Lahiri, P. (2001), 'Empirical best prediction for small area inference with binary data', *Annals of the Institute of Statistical Mathematics* **53**(2), 217–243.

Jiang, J. and Lahiri, P. (2006), 'Mixed model prediction and small area estimation', *Test* **15**(1), 1.

Karlberg, F. (2015), 'Small area estimation for skewed data in the presence of zeroes', *Statistics in Transition new series* **4**(16), 541–562.

Lyu, X. (2020), *saezero: Small Area Estimation under a Zero Inflated Lognormal Model with Correlated Random Area Effects*. R package version 0.1.0.
URL: *https://github.com/XiaodanLyu/saezero*

Marhuenda, Y., Molina, I., Morales, D. and Rao, J. (2017), 'Poverty mapping in small areas under a twofold nested error regression model', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **180**(4), 1111–1136.

Marino, M. F., Ranalli, M. G., Salvati, N., Alfò, M. et al. (2019), 'Semiparametric empirical best prediction for small area estimation of unemployment indicators', *The Annals of Applied Statistics* **13**(2), 1166–1197.

Min, Y. and Agresti, A. (2002), 'Modeling nonnegative data with clumping at zero: a survey', *Journal of the Iranian Statistical Society* **1**(1), 7–33.

Molina, I., Martin, N. et al. (2018), 'Empirical best prediction under a nested error model with log transformation', *The Annals of Statistics* **46**(5), 1961–1993.

Molina, I. and Rao, J. (2010), 'Small area estimation of poverty indicators', *Canadian Journal of Statistics* **38**(3), 369–385.

Nusser, S. M. and Goebel, J. J. (1997), 'The national resources inventory: a long-term multi-resource monitoring programme', *Environmental and Ecological Statistics* **4**(3), 181–204.

Pfeffermann, D., Terryn, B. and Moura, F. A. (2008), 'Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries', *Survey Methodology* **34**(2), 235–249.

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Rao, J. and Molina, I. (2015), *Small Area Estimation*, John Wiley & Sons.

Schall, R. (1991), 'Estimation in generalized linear models with random effects', *Biometrika* **78**(4), 719–727.

Smith, D. D. and Wischmeier, W. H. (1957), 'Factors affecting sheet and rill erosion', *Eos, Transactions American Geophysical Union* **38**(6), 889–896.

Smyth, G. K. (2014), 'Polynomial approximation', *Wiley StatsRef: Statistics Reference Online* .

Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', *Econometrica: journal of the Econometric Society* **26**(1), 24–36.

U.S. Department of Agriculture (2020*a*), 'National Agricultural Statistics Service Cropland Data Layer', Published crop-specific data layer [Online]. USDA-NASS, Washington, DC.
URL: *https://nassgeodata.gmu.edu/CropScape/*

U.S. Department of Agriculture (2020*b*), 'Soil Survey Geographic (SSURGO) Database', Soil Survey Staff, Natural Resources Conservation Service.
URL: *https://sdmdataaccess.sc.egov.usda.gov*

Williams, J. R. and Izaurralde, R. (2006), The APEX model, *in* V. Singh and D. Frevert, eds, 'Watershed Models', Taylor & Francis Group, chapter 18, pp. 437–482.

Wischmeir, W. H. and Smith, D. D. (1965), *Predicting rainfall erosion losses from cropland east of the Rocky Mountains: guide for selection for practices for soil and water conservation*, U.S. Department of Agriculture. Agricultural handbook NO.282.

Zimmermann, T. and Münnich, R. T. (2018), 'Small area estimation with a lognormal mixed model under informative sampling', *Journal of Official Statistics* **34**(2), 523–542.

Figures and tables

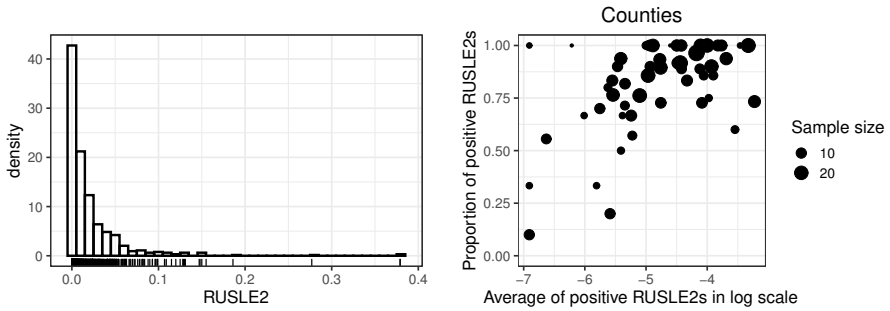

Figure 1 Left: Histogram of CEAP sampled RUSLE2s in South Dakota. Right: Scatterplot of county proportions of positive RUSLE2s against county mean of log RUSLE2s.

Table 1 Average MSE differences ($\times 10^5$) between the alternative predictors (EB(0), EB predictor assuming $\rho = 0$; PI, plug-in predictor; ZI, zero-ignored MMSE predictor; SI, shifted MMSE predictor) and the EB predictor. The associated Monte Carlo margins of error are presented in parentheses.

predictor	EB(0)	PI	ZI	SI
Avg. for $n_i = 5$	1.17 (0.99)	1.20 (0.99)	4.28 (0.96)	71.58 (7.32)
Avg. for $n_i = 10$	0.89 (0.72)	0.90 (0.72)	3.57 (0.70)	78.04 (8.17)
Avg. for $n_i = 20$	0.64 (0.35)	0.63 (0.35)	2.76 (0.36)	71.29 (6.54)

Table 2 Average MSE differences ($\times 10^5$) between the EB(0) predictor and the EB predictor as the true correlation ρ of the random area effects between the positive part and the binary part changes. The associated Monte Carlo margins of error are presented in parentheses.

ρ	-0.9	-0.6	-0.3	0	0.3	0.6	0.9
Avg. for $n_i = 5$	2.43 (0.25)	0.97 (0.18)	0.16 (0.12)	-0.14 (0.08)	-0.05 (0.09)	0.45 (0.13)	1.17 (0.19)
Avg. for $n_i = 10$	2.49 (0.23)	0.79 (0.16)	0.01 (0.13)	-0.14 (0.08)	-0.03 (0.07)	0.27 (0.12)	0.89 (0.16)
Avg. for $n_i = 20$	1.54 (0.13)	0.39 (0.10)	-0.02 (0.06)	-0.08 (0.04)	0.02 (0.04)	0.23 (0.06)	0.64 (0.09)

Table 3 Average biases ($\times 10^5$) and coverage percentages (CP) of nominal 95% normal theory prediction intervals for the MSE estimators (one-step, bootstrap, semi-bootstrap). The associated Monte Carlo margins of error are presented in parentheses.

	One-Step		Bootstrap		Semi-Boot	
	Bias	CP	Bias	CP	Bias	CP
Avg. for $n_i = 5$	-0.66 (0.51)	94.94 (0.33)	0.48 (0.48)	94.54 (0.33)	0.46 (0.53)	95.36 (0.31)
Avg. for $n_i = 10$	-0.96 (0.32)	94.88 (0.32)	-0.00 (0.30)	94.08 (0.33)	-0.14 (0.33)	95.48 (0.30)
Avg. for $n_i = 20$	-0.68 (0.19)	94.47 (0.35)	-0.16 (0.16)	93.78 (0.32)	-0.12 (0.19)	95.37 (0.31)

Table 4 The parameter estimates and associated bootstrap standard errors (SE) for the zero-inflated log-normal model fit to the cropland CEAP RUSLE2 data.

	Positive Part	Binary Part
	Estimate (SE)	Estimate (SE)
logR	2.19 (0.36)	4.94 (0.72)
logK	0.52 (0.23)	
logS	0.49 (0.08)	0.38 (0.21)
is.soybean		0.71 (0.33)
is.sprwht		0.98 (0.52)
Var: county	0.22	0.47
Var: residual	1.23	
Correlation	0.77	

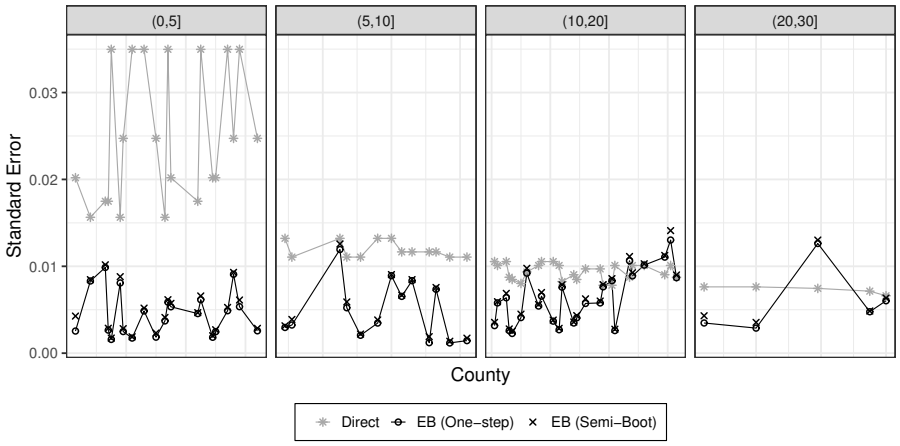

Figure 2 Standard errors of the EB predictor and the direct estimator (sample mean) of mean cropland RUSLE2 South Dakota counties. Standard errors for the EB predictor are square roots of the one-step or semi-boot MSE estimator. The pooled standard error is used for the direct estimator. The comparisons are grouped by sample size labeled on the top.

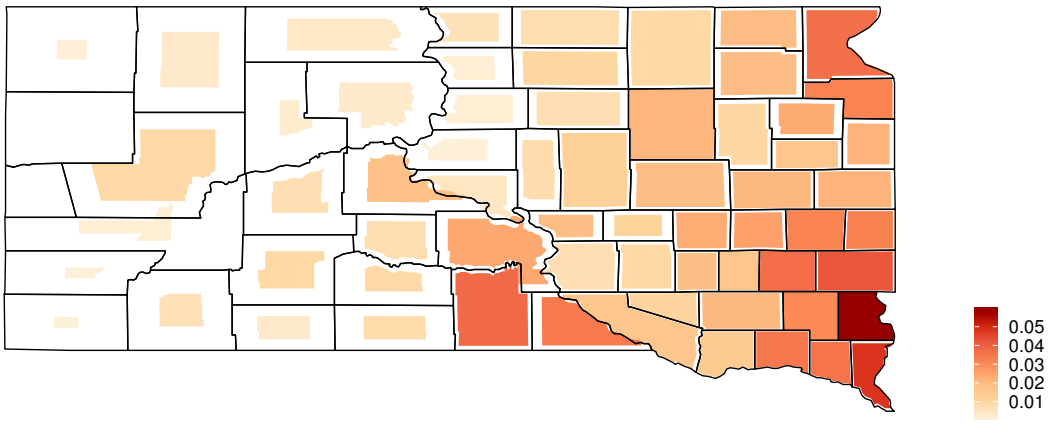

Figure 3 Cartogram of the EB predicted county means of cropland RUSLE2 in South Dakota. Darker shade indicates severer soil sheet and rill erosion. Smaller shrinkage indicates smaller coefficient of variance. This figure appears in color in the electronic version of this article.

