Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Small area prediction of general small area parameters for unit-level count data

by Emily Berg

Release date: January 3, 2024

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2023
Vol. 49, No. 2, pp. 547-568
Statistics Canada, Catalogue No. 12-001-X

Small area prediction of general small area parameters for unit-level count data

Emily Berg[1]

Abstract

We investigate small area prediction of general parameters based on two models for unit-level counts. We construct predictors of parameters, such as quartiles, that may be nonlinear functions of the model response variable. We first develop a procedure to construct empirical best predictors and mean square error estimators of general parameters under a unit-level gamma-Poisson model. We then use a sampling importance resampling algorithm to develop predictors for a generalized linear mixed model (GLMM) with a Poisson response distribution. We compare the two models through simulation and an analysis of data from the Iowa Seat-Belt Use Survey.

Key Words: Poisson; Bootstrap; Small area estimation.

1. Introduction

Small area estimation is the problem of constructing estimators for domains where sample sizes are too small to support reliable direct estimators. The standard approach to small area estimation is to use model-based estimators instead of direct estimators. Model-based estimators garner efficiency gains for small area estimation through restrictions that different areas share common distributional properties and through the incorporation of population-level auxiliary information. Extensive reviews of small area models are available in Rao and Molina (2015), Jiang and Lahiri (2006), and Pfeffermann (2013). More recent reviews include Ghosh (2020) and Molina, Corral and Nguyen (2022). The small area literature has focused heavily on the situation in which the parameter of interest is a small area mean. Many small area parameters are not simple means but are nonlinear functions of the model response variable. Molina and Rao (2010) develops a simulation-based procedure for constructing predictors of small area parameters that may be nonlinear functions of the model response variable. We refer to the types of parameters of interest in Molina and Rao (2010) as "general parameters". Molina, Nandram and Rao (2014), Hobza, Marhuenda and Morales (2020), Rojas-Perilla, Pannier, Schmid and Tzavidis (2020), Marhuenda, Molina, Morales and Rao (2017) and Guadarrama, Molina and Rao (2018) extend Molina and Rao (2010) to Bayesian inference, generalized linear mixed models, data transformations, two-fold models, and complex sampling. We develop predictors of general small area parameters for unit-level count data.

The two primary small area models for count data are (1) the gamma-Poisson model and (2) the Poisson generalized linear mixed model (GLMM). In the context of the area-level model, Reluga, Lombardía and Sperlich (2021) and Boubeta, Lombardía and Morales (2016) develop small area prediction procedures for the gamma-Poisson model and the Poisson GLMM, respectively. We focus on unit-level models. Tzavidis, Ranalli, Salvati, Dreassi and Chambers (2015) develops an M-quantile based procedure for prediction of

1. Emily Berg, Iowa State University. E-mail: emilyb@iastate.edu.

small area means for unit-level count data. As demonstrated in Tzavidis et al. (2015), this procedure is less efficient than model-based methods if the model assumptions hold. Jiang and Lahiri (2006) develops an empirical best predictor of the mean for a unit-level Poisson GLMM. Berg (2022) develops empirical best predictors of the mean under a unit-level gamma-Poisson model. We refer to Berg (2022) for a more complete review of unit-level models and area-level models for count data. Jiang and Lahiri (2006), Tzavidis et al. (2015), and Berg (2022) emphasize prediction of means. We develop procedures that are applicable to nonlinear parameters, such as quantiles.

We propose empirical best predictors of general parameters under two unit-level models. The first is a unit-level gamma-Poisson model. The second is a unit-level Poisson GLMM. We establish a common notation that we use for both models. Let $i = 1, \ldots, D$ index the small areas, and let $j = 1, \ldots, N_i$ index the units in the population for small area i. Let y_{ij} be the observed count for unit j in small area i, where $y_{ij} \in \{0, 1, 2, \ldots\}$. Let $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ be a vector of covariates that does not include an intercept. We consider prediction of a general parameter defined as

$$\theta_i = \theta_i(\mathbf{y}_i) = Q(y_{i1}, \ldots, y_{iN_i}), \tag{1.1}$$

where $Q(\cdot)$ is a real-valued, measurable function, and $\mathbf{y}_i = (y_{i1}, \ldots, y_{iN_i})'$. Common choices of $Q(\cdot)$ are the finite population mean or quantile. Molina and Rao (2010) gives several examples of the function $Q(\cdot)$. Assume y_{ij} is observed only for the elements in the sample. The covariate $\mathbf{x}_{ij}$ is required for every element of the population. In this probabilistic framework, the population U is partitioned into two parts as $U = A \cup R$, where A is the index set of the sample and R is the index set of the non-sampled elements. We partition A and R as $A = \cup_{i=1}^{D} A_i$ and $R = \cup_{i=1}^{D} R_i$, where A_i is the index set of sampled elements for area i, and R_i is the index set of non-sampled elements in area i. Without loss of generality, it can be assumed that $A_i = \{1, \ldots, n_i\}$, and $R_i = \{n_i + 1, \ldots, N_i\}$. With this convention, $\mathbf{y}_i = (\mathbf{y}_{is}', \mathbf{y}_{ir}')'$, where $\mathbf{y}_{is} = (y_{i1}, \ldots, y_{in_i})'$, and $\mathbf{y}_{ir} = (y_{in_i+1}, \ldots, y_{iN_i})'$.

We compare the predictors of θ_i and corresponding MSE estimators based on the gamma-Poisson model to the predictors and MSE estimators based on the Poisson GLMM through simulation. We simulate data from both the Poisson GLMM and the gamma-Poisson model. For each simulation model, we calculate the predictors and MSE estimators based on both the Poisson GLMM and the gamma-Poisson model. This allows us to evaluate the properties of the procedures for situations where the model is correctly specified and under model misspecification.

We illustrate the methods using a subset of data collected in the 2018 Iowa Seat-Belt Use survey. Berg (2022) constructs predictors of county-level means using this same data set. We extend the analysis of Berg (2022) to include predictors of the median and the inter-quartile range (IQR), using the gamma-Poisson model as well as the Poisson GLMM. The data analysis is somewhat contrived to suit our interest in count data. The actual parameters of interest are proportions of belted occupants. In Berg (2023a), we conduct a more extensive analysis of the data that is geared toward the practical needs of the seat-belt use survey. That analysis motivated our interest in developing methodology for small area models for counts. The analysis

in this paper allows us to effectively illustrate the proposed methodology. The methods that we propose are of practical interest beyond the illustrative application. Count data appear frequently in the small area estimation literature (Tzavidis et al., 2015). Applications often benefit from estimates of parameters that are more general than the mean (Molina and Rao, 2010; Hobza et al., 2020). The method that we propose uniquely provides estimates of a broad range of parameters for count data.

Our primary contribution is the development of predictors of nonlinear small area parameters, such as those considered in Molina and Rao (2010), for a situation in which the response variable is a count. The models in this paper are not new, but to our knowledge, the proposed prediction algorithms are novel. Our development of predictors of nonlinear parameters builds on work in Hobza et al. (2020). The methods of Hobza et al. (2020) apply to a general GLMM specification, but their simulations and data analysis focus on the gamma response distribution. We provide detailed steps to construct predictors of nonlinear parameters for count data. Also, Hobza et al. (2020) restricts attention to additive parameters of the form $N_i^{-1}\sum_{j=1}^{N_i} q(y_{ij})$ for a specified function q. The class of nonlinear parameters that we define in (1.1) is more general than the class of additive parameters discussed in Hobza et al. (2020). The class of parameters that we define is broad enough to encompass quantiles. It also includes other parameters of practical importance, such as the small area skewness and kurtotsis. The method of Hobza et al. (2020) is not immediately applicable to estimation of non-additive parameters, such as quantiles, the skewness, and the kurtosis. While the mean provides an indication of the central tendency within a small area, estimates of the quantiles and higher moments provide a more complete picture of the distribution of the characteristic at the small area level.

Our approach has two limitations which are important to assert. First, we assume that the sample design is noninformative for the specified model. If the sampling weights are correlated with the model response variable, after conditioning on model covariates then the design is informative. The methods that we propose will render biased inferences under informative sampling. In ongoing work, Berg and Eideh (2023) extend the proposed methods to a complex sample design. We refer the reader to Parker, Janicki and Holan (2019) for a comprehensive review of small area estimation under informative sampling. A second limitation is that we require the covariate for every unit in the population. In many applications, it may be difficult to satisfy this assumption. If only area-level covariates are available, then area-level models may be preferable.

The rest of this manuscript is organized as follows. In Section 2, we develop empirical best predictors of nonlinear parameters for the unit-level gamma-Poisson model. In Section 3, we develop empirical best predictors of nonlinear parameters for a GLMM with a Poisson response distribution. In Section 4, we compare the two procedures through simulation. In Section 5, we apply both procedures to the seat-belt survey data. We conclude with a discussion highlighting the strengths and weaknesses of the two models Section 6.

2. Unit-level gamma Poisson model and predictor

We define the unit-level gamma-Poisson model as in Berg (2022). Assume

$$y_{ij} \mid \mu_{ij} \overset{\text{ind}}{\sim} \text{Poisson}(\mu_{ij}), \quad i = 1, \ldots, D; \; j = 1, \ldots, N_i, \tag{2.1}$$

where $\mu_{ij} = \lambda_{ij} u_i$, $u_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$, $\lambda_{ij} = g(\mathbf{x}_{ij}'\boldsymbol{\gamma})$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ is a fixed vector of regression coefficients, and $g(\cdot)$ is a specified link function. We let $g(\mathbf{x}_{ij}'\boldsymbol{\gamma}) = \exp(\mathbf{x}_{ij}'\boldsymbol{\gamma})$. The notation $\text{Gamma}(a, b)$ denotes a gamma distribution with shape parameter a and rate parameter b such that $E[u_i] = \alpha/\beta$.

We develop an empirical best predictor of θ_i. As in Berg (2022), the assumptions of the model (2.1) imply that the conditional distribution of u_i given the observed data satisfies,

$$u_i \mid \mathbf{y}_{is} \overset{\text{ind}}{\sim} \text{Gamma}(y_{i.} + \alpha, \beta + \lambda_{i.}), \tag{2.2}$$

where $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$ and $\lambda_{i.} = \sum_{j=1}^{n_i} \lambda_{ij}$. Note that (2.2) holds exactly for any sample size and does not require approximations. The conditional distribution (2.2) is the crux of the development of the empirical best predictors. The conditional distribution (2.2) depends on the unknown α, β, and $\boldsymbol{\gamma}$. To operate with the conditional distribution, we use the maximum likelihood estimators of these fixed parameters. As demonstrated in Berg (2022), the log likelihood for α, β, $\boldsymbol{\gamma}$ is of the form $\ell(\alpha, \beta, \boldsymbol{\gamma}) = \sum_{i=1}^{D} \log\big(L_i(\alpha, \beta, \boldsymbol{\gamma})\big)$, where

$$L_i(\alpha, \beta, \boldsymbol{\gamma}) = \frac{\beta^{\alpha}\left[\prod_{j=1}^{n_i} \lambda_{ij}^{y_{ij}}\right]}{\Gamma(\alpha)\prod_{j=1}^{n_i} y_{ij}!} \frac{\Gamma(y_{i.} + \alpha)}{(\beta + \lambda_{i.})^{y_{i.}+\alpha}}.$$

Define the maximum likelihood estimator by

$$(\hat{\alpha}, \hat{\beta}, \hat{\boldsymbol{\gamma}}')' = \text{argmax}_{(\alpha, \beta, \boldsymbol{\gamma})}\, \ell(\alpha, \beta, \boldsymbol{\gamma}).$$

Berg (2022) discusses the theoretical properties of the maximum likelihood estimator for the gamma-Poisson model.

The known conditional distribution for u_i in (2.2), combined with the maximum likelihood estimator, motivates a computationally simple procedure for predicting θ_i. The procedure is an application of the general method of Molina and Rao (2010) to the gamma-Poisson model (2.1). The best predictor of θ_i under squared error loss is $\tilde{\theta}_i = E\big[\theta_i \mid \alpha, \beta, \boldsymbol{\gamma}, \mathbf{y}_s\big]$. By mutual independence of $\mathbf{y}_{1s}, \ldots, \mathbf{y}_{Ds}$, the best predictor simplifies as $\tilde{\theta}_i = \tilde{\theta}_i(\alpha, \beta, \boldsymbol{\gamma}, \mathbf{y}_{is}) = E\big[\theta_i \mid \alpha, \beta, \boldsymbol{\gamma}, \mathbf{y}_{is}\big]$, where

$$\tilde{\theta}_i = E\big[\theta_i \mid \alpha, \beta, \boldsymbol{\gamma}, \mathbf{y}_{is}\big] = \sum_{y_{in_i+1}=0}^{\infty} \cdots \sum_{y_{iN_i}=0}^{\infty} \theta_i(\mathbf{y}_i)\, f\big(\mathbf{y}_{ir} \mid \mathbf{y}_{is}; \alpha, \beta, \boldsymbol{\gamma}\big),$$

and

$$f\big(\mathbf{y}_{ir} \mid \mathbf{y}_{is}; \alpha, \beta, \boldsymbol{\gamma}\big) = \left(\frac{\prod_{j=n_i+1}^{N_i} \lambda_{ij}^{y_{ij}}}{\prod_{j=n_i+1}^{N_i} y_{ij}!}\right)\left(\frac{\Gamma\left(\sum_{j=1}^{N_i} y_{ij} + \alpha\right)}{\Gamma\left(\sum_{j=1}^{n_i} y_{ij} + \alpha\right)}\right)\left(\frac{(\beta + \lambda_{i.})^{y_{i.}+\alpha}}{\left(\beta + \sum_{j=1}^{N_i} \lambda_{ij}\right)^{\sum_{j=1}^{N_i} y_{ij}+\alpha}}\right).$$

The notation $\tilde{\theta}_i(\alpha, \beta, \boldsymbol{\gamma}, \mathbf{y}_{is})$ emphasizes dependence of the best predictor on the unknown $\alpha, \beta,$ and $\boldsymbol{\gamma}$. An empirical best predictor is obtained by substitution of the unknown parameters defining the best predictor with the maximum likelihood estimators. That is, the empirical best predictor is defined as

$$\hat{\theta}_i = E\left[\theta_i \,|\, \hat{\alpha}, \hat{\beta}, \hat{\boldsymbol{\gamma}}, \mathbf{y}_{is}\right] = \sum_{y_{in_i+1}=0}^{\infty} \cdots \sum_{y_{iN_i}=0}^{\infty} \theta_i(\mathbf{y}_i)\, f\left(\mathbf{y}_{ir} \,|\, \mathbf{y}_{is}; \hat{\alpha}, \hat{\beta}, \hat{\boldsymbol{\gamma}}\right). \tag{2.3}$$

The infinite sum defining the empirical best predictor is analytically intractable. We define a Monte Carlo (MC) approximation for the empirical best predictor of θ_i, as in Molina and Rao (2010).

We define L simulated populations. For $\ell = 1, \ldots, L,$ we set $y_{ij}^{(\ell)} = y_{ij}$ for $j = 1, \ldots, n_i$. For non-sampled elements, $j = n_i + 1, \ldots, N_i,$ we generate $y_{ij}^{(\ell)}$ as $y_{ij}^{(\ell)} \,|\, \mu_{ij}^{(\ell)} \sim \text{Poisson}\left(\mu_{ij}^{(\ell)}\right),$ where

$$\begin{aligned} \mu_{ij}^{(\ell)} &= \exp\left(\mathbf{x}_{ij}'\hat{\boldsymbol{\gamma}}\right) u_i^{(\ell)}, \\ u_i^{(\ell)} &\sim \text{Gamma}\left(y_{i.} + \hat{\alpha}, \hat{\beta} + \hat{\lambda}_{i.}\right), \end{aligned}$$

and $\hat{\lambda}_{i.} = \sum_{j=1}^{n_i} \exp\left(\mathbf{x}_{ij}'\hat{\boldsymbol{\gamma}}\right)$. An MC approximation for the empirical best predictor of θ_i is then

$$\hat{\theta}_{i,L} = \hat{\theta}_{i,L}\left(\hat{\boldsymbol{\gamma}}, \hat{\alpha}, \hat{\beta}, \mathbf{y}_{is}\right) = L^{-1} \sum_{\ell=1}^{L} \theta_i^{(\ell)}, \tag{2.4}$$

where $\theta_i^{(\ell)} = Q\left(y_{i1}^{(\ell)}, \ldots, y_{iN_i}^{(\ell)}\right)$. We express the predictor $\hat{\theta}_{i,L}$ as a function of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\boldsymbol{\gamma}}$, and $\mathbf{y}_{is}$ to emphasize dependence of the predictor on the parameter estimators and the observed counts for the area.

2.1 MSE Estimation

We use the bootstrap method of González-Manteiga, Lombardía, Molina, Morales and Santamaría (2007) and Molina and Rao (2010) to estimate the MSE of $\hat{\theta}_{i,L}$. For $b = 1, \ldots, B,$ we repeat the following steps:

1. For $i = 1, \ldots, D,$ generate $\{y_{i1}^{*(b)}, \ldots, y_{iN_i}^{*(b)}\}$ from the model in (2.1) with parameters equal to the maximum likelihood estimate $(\hat{\alpha}, \hat{\beta}, \hat{\boldsymbol{\gamma}})'$. Define a bootstrap version of the population parameter by $\theta_i^{*(b)} = Q\left(y_{i1}^{*(b)}, \ldots, y_{iN_i}^{*(b)}\right)$. Let $\mathbf{y}_{is}^{*(b)} = (y_{i1}^{*(b)}, \ldots, y_{in_i}^{*(b)})'$. The bootstrap sample is then $\{\mathbf{y}_{is}^{*(b)} : i = 1, \ldots, D\}$.
2. Use the bootstrap sample, $\left\{\mathbf{y}_{is}^{*(b)} : i = 1, \ldots, D\right\}$, to obtain a maximum likelihood estimator denoted as $\left(\hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)}, (\hat{\boldsymbol{\gamma}}^{*(b)})'\right)'$. Specifically, $\left(\hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)}, (\hat{\boldsymbol{\gamma}}^{*(b)})'\right)'$ satisfies

$$\left(\hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)}, \left(\hat{\boldsymbol{\gamma}}^{*(b)}\right)'\right)' = \text{argmax}_{(\alpha,\beta,\boldsymbol{\gamma})}\, \ell^{*(b)}(\alpha, \beta, \boldsymbol{\gamma}),$$

where

$$\ell^{*(b)}(\alpha, \beta, \boldsymbol{\gamma}) = \sum_{i=1}^{D} \log\left(L_i^{*(b)}(\alpha, \beta, \boldsymbol{\gamma})\right),$$

$$L_i^{*(b)}(\alpha, \beta, \boldsymbol{\gamma}) = \frac{\beta^{\alpha}\left[\prod_{j=1}^{n_i} \lambda_{ij}^{y_{ij}^{*(b)}}\right]}{\Gamma(\alpha)\prod_{j=1}^{n_i} y_{ij}^{*(b)}!} \frac{\Gamma\left(y_{i.}^{*(b)}+\alpha\right)}{(\beta+\lambda_{i.})^{y_{i.}^{*(b)}+\alpha}},$$

and $y_{i.}^{*(b)} = \sum_{j=1}^{n_i} y_{ij}^{*(b)}$.

3. Construct the MC approximation for the empirical best predictor using the bootstrap sample and the bootstrap maximum likelihood estimator. Denote the bootstrap version of the empirical best predictor by $\hat{\theta}_{i,L}^{*(b)} = \hat{\theta}_i\left(\hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)}, \hat{\boldsymbol{\gamma}}^{*(b)}, \mathbf{y}_{is}^{*(b)}\right)$. We construct $\hat{\theta}_{i,L}^{*(b)}$ as follows. We define L simulated populations. For $\ell = 1, \ldots, L$, we set $y_{ij}^{(\ell)} = y_{ij}^{*(b)}$ for $j = 1, \ldots, n_i$. For non-sampled elements, $j = n_i + 1, \ldots, N_i$, we generate $y_{ij}^{(\ell *b)}$ as $y_{ij}^{(\ell *b)} \mid \mu_{ij}^{(\ell *b)} \sim \text{Poisson}\left(\mu_{ij}^{(\ell *b)}\right)$, where

$$\begin{aligned} \mu_{ij}^{(\ell *b)} &= \exp(\mathbf{x}_{ij}'\hat{\boldsymbol{\gamma}}^{*(b)})\, u_i^{(\ell *b)}, \\ u_i^{(\ell *b)} &\sim \text{Gamma}\,(y_{i.}^{*(b)} + \hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)} + \hat{\lambda}_{i.}^{*(b)}), \end{aligned}$$

$y_{i.}^{*(b)} = \sum_{j=1}^{n_i} y_{ij}^{*(b)}$, and $\hat{\lambda}_{i.}^{*(b)} = \sum_{j=1}^{n_i} \exp(\mathbf{x}_{ij}'\hat{\boldsymbol{\gamma}}^{*(b)})$. We then define $\hat{\theta}_{i,L}^{*(b)} = \hat{\theta}_{i,L}(\hat{\boldsymbol{\gamma}}^{*(b)}, \hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)}, \mathbf{y}_{is}^{*(b)}) = L^{-1}\sum_{\ell=1}^{L} \theta_i^{(\ell *b)}$, where $\theta_i^{(\ell *b)} = Q(y_{i1}^{(\ell *b)}, \ldots, y_{iN_i}^{(\ell *b)})$.

Define a bootstrap estimator of the MSE by

$$\widehat{\text{MSE}}_i = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{i,L}^{*(b)} - \hat{\theta}_i^{*(b)}\right)^2. \tag{2.5}$$

3. Poisson GLMM

We next define an empirical best prediction procedure for a GLMM with a Poisson response distribution. The Poisson GLMM assumes that

$$y_{ij} \mid \mu_{ij} \overset{\text{ind}}{\sim} \text{Poisson}(\mu_{ij}), \quad i = 1, \ldots, D;\ j = 1, \ldots, N_i, \tag{3.1}$$

$$h(\mu_{ij}) = \beta_0 + \mathbf{x}_{ij}'\boldsymbol{\beta}_1 + b_i,$$

and $b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2)$ for $i = 1, \ldots, D$. The function $h(\mu_{ij})$ is a specified link function. We assume that $h(\mu_{ij}) = \log(\mu_{ij})$.

We estimate the parameters of the Poisson GLMM using the method of Schall (1991). This method is applicable to a very general GLMM. The method of Schall (1991) is used in the context of binomial data in González-Manteiga et al. (2007). The supplementary material of Berg (2022) describes the steps of the Schall (1991) procedure for the specific Poisson GLMM. Because the method has the form of an iteratively reweighted least squares (IRLS) procedure, we refer to the algorithm of Schall (1991) as the IRLS algorithm. Let $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}_1$, $\hat{b}_i$ and $\hat{\sigma}_b^2$ be the estimators and predictors obtained upon completion of the IRLS algorithm.

Remark: The `R` function `glmer` is a widely used alternative to the IRLS algorithm. We emphasize the IRLS algorithm in the main document because the IRLS algorithm is reproducible in programming languages other than `R`. We present results using `glmer` in Section 2 of the supplementary material (Berg, 2023b).

3.1 Empirical best predictor for Poisson GLMM

The best predictor of θ_i under squared error loss is

$$E\left[\theta_i \,|\, \mathbf{y}_{is}; \beta_0, \boldsymbol{\beta}_1, \sigma_b^2\right] = \sum_{y_{in_i+1}=0}^{\infty} \cdots \sum_{y_{iN_i}=0}^{\infty} \theta_i(\mathbf{y}_i)\, f\left(\mathbf{y}_{ir} \,|\, \mathbf{y}_{is}\right),$$

where

$$f\left(\mathbf{y}_{ir} \,|\, \mathbf{y}_{is}\right) = \int_{-\infty}^{\infty} \left[\prod_{j=n_i+1}^{N_i} \mu_{ij}^{y_{ij}} \exp(-\mu_{ij}) \big/ y_{ij}!\right] f\left(b_i \,|\, \mathbf{y}_{is}\right) db_i,$$

$$f\left(b_i \,|\, \mathbf{y}_{is}\right) = \frac{\left[\prod_{j=1}^{n_i} \mu_{ij}^{y_{ij}} \exp(-\mu_{ij}) \big/ y_{ij}!\right] \phi\left(b_i/\sigma_b\right)/\sigma_b}{\int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \mu_{ij}^{y_{ij}} \exp(-\mu_{ij}) \big/ y_{ij}!\right] \phi\left(b_i/\sigma_b\right)/\sigma_b db_i},$$

and ϕ is the pdf of a standard normal distribution. The empirical best predictor is then $E\left[\theta_i \,|\, \mathbf{y}_{is}; \hat{\beta}_0, \hat{\boldsymbol{\beta}}_1, \hat{\sigma}_b^2\right]$.

The conditional distribution of b_i given the data does not have a known form for the GLMM. We therefore require a Monte Carlo procedure to approximate this conditional distribution. We use a method called sampling importance resampling (SIR) to obtain a Monte Carlo approximation for the empirical best predictor of (1.1) under the assumptions of the model (3.1). The SIR algorithm is traditionally used to sample from posterior distributions in a Bayesian context (Smith and Gelfand, 1992). We use SIR for the purpose of obtaining an MC approximation for the empirical best predictor. The SIR algorithm involves simulating from a proposal distribution and then accepting a proposed value with probability proportional to the ratio of the target and proposal distributions. The SIR algorithm is a general algorithm, and the details of implementation depend on the context. We describe how we implement the SIR algorithm for the specific Poisson GLMM in steps 1-2 below.

For $l = 1, \ldots, L,$ repeat the following steps:

1. Generate $b_i^{(\ell,1)}, \ldots, b_i^{(\ell,T)} \overset{\text{iid}}{\sim} N(\hat{b}_i, \hat{\sigma}_b^2)$ for $i = 1, \ldots, D.$
2. Define

$$p_i^{(\ell,t)} = \exp\left(\ell_{1i}^{(\ell,t)} - \frac{1}{T}\sum_{t=1}^{T} \ell_{1i}^{(\ell,t)}\right) \left\{\frac{\phi\left(b_i^{(\ell,t)}/\hat{\sigma}_b\right)}{\phi\left((b_i^{(\ell,t)} - \hat{b}_i)/\hat{\sigma}_b\right)}\right\},$$

where

$$\ell_{1i}^{(\ell,t)} = \log\left(\left\{\prod_{j=1}^{n_i} \frac{\mu_{ij}^{(\ell,t)} \exp(-\mu_{ij}^{(\ell,t)})}{y_{ij}!}\right\}\right),$$

and $\log(\mu_{ij}^{(\ell,t)}) = \hat{\beta}_0 + \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}}_1 + b_i^{(\ell,t)}.$ Set

$$\tilde{p}_i^{(\ell,t)} = \frac{p_i^{(\ell,t)}}{\sum_{t=1}^{T} p_i^{(\ell,t)}}.$$

Set $b_i^{(\ell)} = b_i^{(\ell,t)}$ with probability $\tilde{p}_i^{(\ell,t)}$ for $i = 1, \ldots, D$.

3. Generate $y_{ij}^{(\ell)} \sim \text{Poisson}\left(\mu_{ij}^{(\ell)}\right)$, where $\log\left(\mu_{ij}^{(\ell)}\right) = \hat{\boldsymbol{\beta}}_0 + \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}}_1 + b_i^{(\ell)}$ for $i = 1, \ldots, D$, and $j = n_i + 1, \ldots, N_i$. Set $y_{ij}^{(\ell)} = y_{ij}$ for $j = 1, \ldots, n_i$.
4. Define $\theta_i^{(\ell)\text{GLMM}} = Q\left(y_{i1}^{(\ell)}, \ldots, y_{iN_i}^{(\ell)}\right)$.

Finally, define the predictor of θ_i by

$$\hat{\theta}_i^{\text{GLMM}} = \hat{\theta}_i^{\text{GLMM}}\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1, \hat{\sigma}_b^2, \mathbf{y}_{is}\right) = L^{-1}\sum_{\ell=1}^{L}\theta_i^{(\ell)\text{GLMM}}. \tag{3.2}$$

3.2 Bootstrap MSE estimator for Poisson GLMM

We use the parametric bootstrap for MSE estimation. The bootstrap procedure is essentially that of Molina and Rao (2010) and González-Manteiga et al. (2007), applied to the Poisson GLMM. For $b = 1, \ldots, B$, repeat the following steps:

1. Generate a bootstrap population from the model (3.1), with parameters equal to the estimated parameters. Specifically, for $i = 1, \ldots, D$, and $j = 1, \ldots, N_i$, generate

$$y_{ij}^{(b)} \,\big|\, \mu_{ij}^{(b)} \sim \text{Poisson}(\mu_{ij}^{(b)}),$$

where $\log(\mu_{ij}^{(b)}) = \hat{\beta}_0 + \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}}_1 + b_i^{(b)}$, and $b_1^{(b)}, \ldots, b_D^{(b)} \overset{\text{iid}}{\sim} N(0, \hat{\sigma}_b^2)$. Define the bootstrap version of the population parameter by $\theta_i^{*(b)} = Q(y_{i1}^{*(b)}, \ldots, y_{iN_i}^{*(b)})$.

2. Let $\mathbf{y}_{is}^{*(b)} = (y_{i1}^{*(b)}, \ldots, y_{in_i}^{*(b)})$ denote the generated values for the index set in the sample. We call $\{\mathbf{y}_{1s}^{*(b)}, \ldots, \mathbf{y}_{Ds}^{*(b)}\}$ the bootstrap sample.
3. Apply the IRLS method of Schall (1991), described above, to the bootstrap sample, $\{\mathbf{y}_{1s}^{*(b)}, \ldots, \mathbf{y}_{Ds}^{*(b)}\}$, to obtain bootstrap versions of the parameter estimates. Denote the estimates obtained from the bootstrap sample by $(\hat{\beta}_0^{*(b)}, (\hat{\boldsymbol{\beta}}_1^{*(b)})', \hat{\sigma}_b^{*(b)2})'$.
4. Implement the procedure of Section 3.1 with the bootstrap sample and the bootstrap estimates $(\hat{\beta}_0^{*(b)}, (\hat{\boldsymbol{\beta}}_1^{*(b)})', \hat{\sigma}_b^{*(b)2})'$ to obtain a bootstrap version of the predictor. The bootstrap version of the predictor is $\hat{\theta}_i^{*(b)\text{GLMM}} = \hat{\theta}_i^{\text{GLMM}}(\hat{\beta}_0^{*(b)}, \hat{\boldsymbol{\beta}}_1^{*(b)}, \hat{\sigma}_b^{*(b)2}, \mathbf{y}_{is}^{*(b)})$.

Define the bootstrap MSE estimator by

$$\widehat{\text{MSE}}_i^{\text{GLMM}} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_i^{*(b)\text{GLMM}} - \theta_i^{*(b)}\right)^2. \tag{3.3}$$

4. Simulations

The simulation study has primary and secondary objectives. The primary objectives of the simulation study are two-fold. The first is to evaluate the performance of the small area predictors based on the gamma-Poisson model relative to the small area predictors based on the Poisson-GLMM. The second goal is to evaluate the quality of the proposed bootstrap MSE estimators. The two secondary goals of the simulation study are (1) to evaluate the computational time of the alternative procedures, and (2) to assess the choice of T for the SIR algorithm. We present output related to the primary objectives of the simulation in this main document. We relegate further discussion of the secondary objectives of the simulation to the supplementary material (Berg, 2023b). We refer the reader to Berg (2022) for a study of the properties of the estimators of the fixed parameters.

We generate data from both the gamma-Poisson model defined in (2.1) and from the unit-level Poisson GLMM defined in (3.1). For each simulation model, we calculate predictors based on the gamma-Poisson model and the GLMM. This permits an evaluation of the procedures under model misspecification. We simulate a univariate covariate as $x_{ij} \overset{\text{iid}}{\sim} N(0.5,1)$ for $i=1,\ldots,D,$ and $j=1,\ldots,N_i,$ where $N_i=100$ for $i=1,\ldots,D,$ and $D=100.$ The covariate is held fixed across simulation runs. We simulate a population from either the gamma-Poisson model (2.1) or the GLMM (3.1). For the gamma-Poisson model, we set $\gamma=1$ and $\beta=2.$ We use two values of α for the gamma-Poisson model. We first use $\alpha=5.$ We then generate a more skewed distribution by setting $\alpha=$ 0.5. For the GLMM, we set $\beta_0=$ 0.5 and $\beta_1=$ 0.5. We use 0.5 and 1.5 as the two values for σ_b^2 for the GLMM. The combination of two model forms (gamma-Poisson and GLMM) with two values for each of σ_b^2 and α results in a total of four data generating models. We select a simple random sample from each area with a common sampling rate of 5%. The use of a constant sampling rate is fairly unrealistic but is chosen intentionally for two reasons. The first is simplicity. The second is to construct a situation with sample sizes that are small enough to reflect the challenges in real small area problems. We construct predictors of three small area parameters of interest: the area mean, the area median, and the area inter-quartile range (IQR). The area IQR is defined as the difference between the 75-percentile and the 25-percentile for the area.

We construct the two main proposed predictors of each parameter. We use Gam-Pois to denote the empirical best predictor for the gamma-Poisson model. The Gam-Pois predictor is defined in (2.4), where it is denoted as $\hat{\theta}_{i,L}.$ We use GLMM to denote the empirical best predictor for the Poisson-GLMM. The GLMM predictor is denoted $\hat{\theta}_i^{\text{GLMM}}$ where it is defined in (3.2). When implementing the GLMM procedure, we use $T=200$ for the Monte Carlo SIR algorithm. The choice of T is discussed in the supplementary material (Berg, 2023b). For both the GLMM and Gam-Pois predictors, we use $L=$ 1,000. The choice of L is based on a comparison of $L=100$ to $L=$ 1,000. The results for $L=100$ are presented in Section 5 of the supplementary material (Berg, 2023b).

We also compute a direct estimator of each parameter. The direct estimator of the mean is the sample mean for the area. The direct estimator of the median is the sample median. The direct estimator of the IQR is calculated as the difference between the sample 75-percentile and the sample 25-percentile for the area.

For the GLMM, we further define plug-in (PI) predictors. The PI predictor of the mean is defined as

$$\hat{\bar{y}}_i^{\mathrm{PI}} = \frac{1}{N_i}\sum_{j=1}^{N_i}\hat{y}_{ij}^{\mathrm{PI}}, \tag{4.1}$$

where $\hat{y}_{ij}^{\mathrm{PI}} = y_{ij}$ for $j = 1, \ldots, n_i$, $\hat{y}_{ij}^{\mathrm{PI}} = \exp\left(\hat{\beta}_0 + x_{ij}\hat{\beta}_1 + \hat{b}_i\right)$ for $j = n_i + 1, \ldots, N_i$, and $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{b}_i\right)$ is obtained from the IRLS algorithm used for the GLMM model (3.1). We define a PI predictor of the median as the median of $\left\{\hat{y}_{ij}^{\mathrm{PI}} : j = 1, \ldots, N_i\right\}$. We define the PI predictor of the IQR as the difference between the 75-percentile of $\left\{\hat{y}_{ij}^{\mathrm{PI}} : j = 1, \ldots, N_i\right\}$ and the 25-percentile of $\left\{\hat{y}_{ij}^{\mathrm{PI}} : j = 1, \ldots, N_i\right\}$. A PI predictor of the form (4.1) is compared to the M-quantile predictor in Tzavidis et al. (2015).

For the mean, we also calculate the closed form expression for the empirical best predictor based on the gamma-Poisson model. This predictor is defined in Berg (2022). We refer to the predictor of Berg (2022) as Gam-Pois-Alt in the tables below.

The procedures for the median and the IQR require calculating percentiles of a set of numbers. Many procedures to calculate percentiles exist. We calculate all percentiles using the default method in the `R` function `quantile`.

4.1 Comparison of efficiency of alternative predictors

We compare the predictors using two criteria. To define the criteria, let $\hat{\theta}_i^{(m)}$ and $\theta_i^{(m)}$, respectively, denote a predictor of θ_i and corresponding population parameter obtained in MC simulation m, where $m = 1, \ldots, M$. The first criterion is the average relative root mean square error defined by

$$\%\mathrm{RRMSE} = 100\frac{1}{D}\sum_{i=1}^{D}\mathrm{RMSE}_i, \tag{4.2}$$

where $\mathrm{RMSE}_i = \sqrt{\mathrm{MSE}_{\mathrm{MC}i}}\left[\frac{1}{M}\sum_{m=1}^{M}\theta_i^{(m)}\right]^{-1}$, and

$$\mathrm{MSE}_{\mathrm{MC}i} = M^{-1}\sum_{m=1}^{M}\left(\hat{\theta}_i^{(m)} - \theta_i^{(m)}\right)^2. \tag{4.3}$$

The second criterion is the percent average absolute relative bias defined by

$$\%\mathrm{RB} = 100\frac{1}{D}\sum_{i=1}^{D}\mathrm{RB}_i, \tag{4.4}$$

where $\mathrm{RB}_i = M^{-1}\left|\sum_{m=1}^{M}(\hat{\theta}_i^{(m)} - \theta_i^{(m)})\right|\left[M^{-1}\sum_{m=1}^{M}\theta_i^{(m)}\right]^{-1}$. We report the %RRMSE and %RB. We use a Monte Carlo (MC) simulation size of $M = 500$.

We first simulate data from the gamma-Poisson model defined in (2.1). Table 4.1 contains the %RRMSE and %RB of the alternative predictors when the data are generated from the model (2.1). The direct estimator is very inefficient for this simulation because the area sample size is only 5. For this configuration, the gamma-Poisson model is the true model, so it is not surprising that the Gam-Pois-Alt predictor has the smallest %RRMSE for the mean. Likewise, the Gam-Pois predictor is the most efficient predictor for the

median and IQR. Unlike the Gam-Pois-Alt predictor, the Gam-Pois predictor is constructed from L simulated samples and is therefore subject to an extra layer of Monte Carlo variability. The difference between the Gam-Pois predictor and the Gam-Pois-Alt predictor shows the effect of the Monte Carlo error from the L simulated samples on the efficiency of the Gam-Pois predictor. Even though the GLMM is mis-specified, the loss of efficiency from using the GLMM for the IQR and the median is much less than the loss from using the direct estimator. The PI predictor is a poor predictor of the median and the IQR. For the median and the IQR, the %RRMSE of the PI predictor exceeds the %RRMSE of the GLMM and Gam-Pois predictors. The bias makes an important contribution to the MSE for the PI predictor of the median and the IQR. The bias is negligible for the Gam-Pois and GLMM predictors, indicating that for these predictors, the contribution from the variance to the MSE is more important than the contribution from the bias.

Second, we simulate data from the GLMM defined in (3.1). Table 4.2 contains the %RRMSE and %RB when the true model is the Poisson GLMM defined in (3.1). The direct estimator is inefficient, compared to the model-based predictors. The PI predictor is efficient for the mean when $\sigma_b^2 = 0.5$, and the GLMM predictor has the smallest %RRMSE when $\sigma_b^2 = 1.5$. The GLMM predictor has smaller %RRMSE than the Gam-Pois and Gam-Pois-Alt predictors. This is not surprising because the GLMM is the true model for the simulation used to construct Table 4.2. Even though the Gam-Pois model is incorrectly specified, the loss of efficiency from using the Gam-Pois predictor is much smaller than the loss of efficiency from using the direct estimator, relative to the predictors based on the GLMM. For the GLMM and Gam-Pois predictors, the %RB is negligible compared to the %RRMSE, indicating that the contribution from the variance to the overall MSE of the predictor is more important than the contribution from the bias. The PI predictor is less efficient than the Gam-Pois predictor or the GLMM predictor for the median and for the IQR. The PI predictor has a severe bias for predicting the median and the IQR.

The comparison of predictors leads to three main conclusions. First, the relative efficiencies of the predictors depend on the data generating model. If the gamma-Poisson model is true, then the Gam-Pois-Alt predictor is most efficient for the mean, and the Gam-Pois predictor is most efficient for nonlinear parameters. When the GLMM is true, the PI/GLMM predictors are most efficient. Second, the Gam-Pois and GLMM predictors appear to have reasonable efficiency, even under model mis-specification. When the Gam-Pois model is correctly specified, the loss of efficiency from incorrect use of the GLMM predictor is slight. Similarly, the ratio of the %RRMSE of the Gam-Pois predictor to the %RRMSE of the GLMM predictor when the GLMM is true is usually about 1.01. The loss of efficiency from use of the Gam-Pois predictor when the GLMM is true is greater than the loss from the use of the GLMM when the Gam-Pois is true, but not by much. The third conclusion concerns the properties of the PI predictor. The PI predictor is not an estimator of an optimal predictor but nonetheless has good properties for predicting the mean in our simulations. For predicting the median and the IQR, the PI predictor has a substantial enough bias that the PI predictor is less efficient than the GLMM or Gam-Pois predictor. Given our interest in a broad range of parameters, we prefer the GLMM predictor over the PI predictor.

Table 4.1
%RB and %RRMSE of alternative predictors when the true model is the gamma-Poisson model and $L =$ 1,000.

	σ_b^2	Mean		Med.		IQR	
		%RRMSE	%RB	%RRMSE	%RB	%RRMSE	%RB
Gam-Pois	5.000	17.362	0.667	20.263	0.798	19.406	0.764
Gam-Pois-Alt	5.000	17.355	0.670				
GLMM	5.000	17.452	0.711	20.340	0.877	19.473	0.782
PI	5.000	17.450	0.718	20.837	3.209	23.004	10.676
Direct	5.000	60.966	2.026	79.232	15.558	78.220	16.629
Gam-Pois	0.500	55.451	1.882	109.729	3.929	64.552	2.145
Gam-Pois-Alt	0.500	55.431	1.888				
GLMM	0.500	56.584	2.850	110.644	4.558	65.023	2.332
PI	0.500	56.508	2.222	121.632	45.899	79.661	26.730
Direct	0.500	106.654	3.906	237.870	38.205	127.712	17.089

Notes: Gam-Pois = gamma-Poisson; GLMM = Generalized linear mixed model; IQR = Inter-quartile range; PI = Plug-in; RB = Relative biases; RRMSE = Relative root mean square error.

Table 4.2
%RB and %RRMSE of alternative predictors when the true model is the Poisson-GLMM model and $L =$ 1,000.

	σ_b^2	Mean		Med.		IQR	
		%RRMSE	%RB	%RRMSE	%RB	%RRMSE	%RB
Gam-Pois	0.500	24.571	0.816	29.014	1.014	25.128	0.977
Gam-Pois-Alt	0.500	24.555	0.810				
GLMM	0.500	24.249	0.847	28.688	1.036	24.900	0.967
PI	0.500	24.239	0.820	29.306	4.696	42.980	34.347
Direct	0.500	38.820	1.399	53.578	4.722	64.172	24.574
Gam-Pois	1.500	19.995	0.624	22.728	0.758	24.107	0.825
Gam-Pois-Alt	1.500	19.971	0.627				
GLMM	1.500	19.781	0.642	22.526	0.747	23.940	0.827
PI	1.500	19.788	0.609	23.140	3.220	35.215	23.591
Direct	1.500	52.486	1.938	67.632	4.636	95.293	24.036

Notes: Gam-Pois = gamma-Poisson; GLMM = Generalized linear mixed model; IQR = Inter-quartile range; PI = Plug-in; RB = Relative biases; RRMSE = Relative root mean square error.

4.2 Properties of bootstrap MSE estimator

We next consider the properties of the MSE estimators for the gamma-Poisson and GLMM models. The MSE estimator is defined in (2.5) for the gamma-Poisson model and in (3.3) for the GLMM. We calculate both MSE estimates under each data generating model. This allows us to evaluate the properties of the MSE estimates under model misspecification and when the model is correctly specified. The bootstrap sample size is $B = 200$. The choice of $B = 200$ follows from a recommendation in Hobza et al. (2020) that the bootstrap sample size be at least 200. To reduce the computational requirements, we use $L = 100$ for the simulations in this section. To evaluate the MSE estimators, we conduct a simulation with $M = 250$ simulated samples using the same x_{ij} used for the first simulation. We calculate the relative bias of the MSE estimator as well as empirical coverages of normal theory 95% prediction intervals.

We first define the relative bias of the MSE estimator. Let $\widehat{\text{MSE}}_i^{(m)}$ denote an MSE estimate from simulated sample m for $m = 1, \ldots, M$, where $M = 250$. We then define the % relative bias of the MSE estimator for area i by

$$\text{RB}_i = 100\left(\frac{\frac{1}{M}\sum_{m=1}^{M}\widehat{\text{MSE}}_i^{(m)}}{\text{MSE}_{\text{MC}i}} - 1\right)$$

where $\text{MSE}_{\text{MC}i}$ is defined in (4.3) and is based a separate simulation with $M =$ 5,000. We use the output from a separate simulation with $M =$ 5,000 to define the denominator of the RB_i to reduce the variance of RB_i.

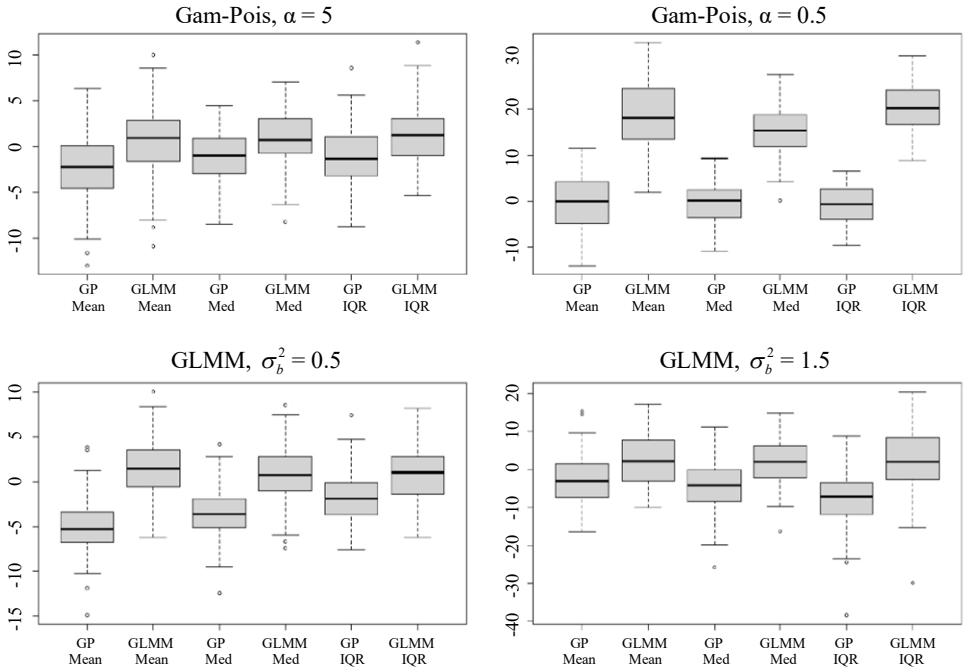
Figure 4.1 contains box-plots of the relative biases for the four simulation configurations and the three parameters. The relative biases of the bootstrap MSE estimator (depicted in Figure 4.1) depend on the simulation model and the parameters. We first consider the Gam-Pois model with $\alpha = 5$. For this configuration, the Gam-Pois MSE estimator has a slight negative bias, but the relative bias is usually between -10% and 10%. The GLMM MSE estimator is nearly unbiased for the MSE of the GLMM predictor in the sense that the relative biases for the GLMM MSE estimator are symmetric around zero and are usually between -10% and 10%. It is interesting that the GLMM MSE estimator performs well for this simulation configuration because the GLMM model is incorrectly specified. We next consider the Gam-Pois model with $\alpha =$ 0.5. The Gam-Pois MSE estimator remains nearly unbiased, in the sense that the median relative bias is close to zero and the relative biases are usually below 10% in absolute value. The GLMM MSE estimator for the GLMM predictor tends to have a positive bias under the Gam-Pois model with $\alpha =$ 0.5. We next consider the GLMM simulation model with $\sigma_b^2 =$ 0.5. The GLMM MSE estimator for the GLMM predictor is nearly unbiased, with most relative biases between -10% and 10%. The Gam-Pois MSE estimator tends to underestimate the MSE of the Gam-Pois predictor for the GLMM simulation configuration with $\sigma_b^2 =$ 0.5. The relative bias of the Gam-Pois MSE estimator is only about -5% for this configuration. Increasing σ_b^2 to 1.5 has little impact on the properties of the GLMM MSE estimator for the GLMM predictors. For the GLMM simulation model with $\sigma_b^2 =$ 1.5, the Gam-Pois MSE estimator can exhibit extreme values.

We define the relative root mean square error of the MSE estimator as

$$\text{RRMSEMSE}_i = \frac{\sqrt{M^{-1}\sum_{m=1}^{M}\left(\widehat{\text{MSE}}_i^{(m)} - \text{MSE}_{\text{MC}i}\right)^2}}{\text{MSE}_{\text{MC}i}}.$$

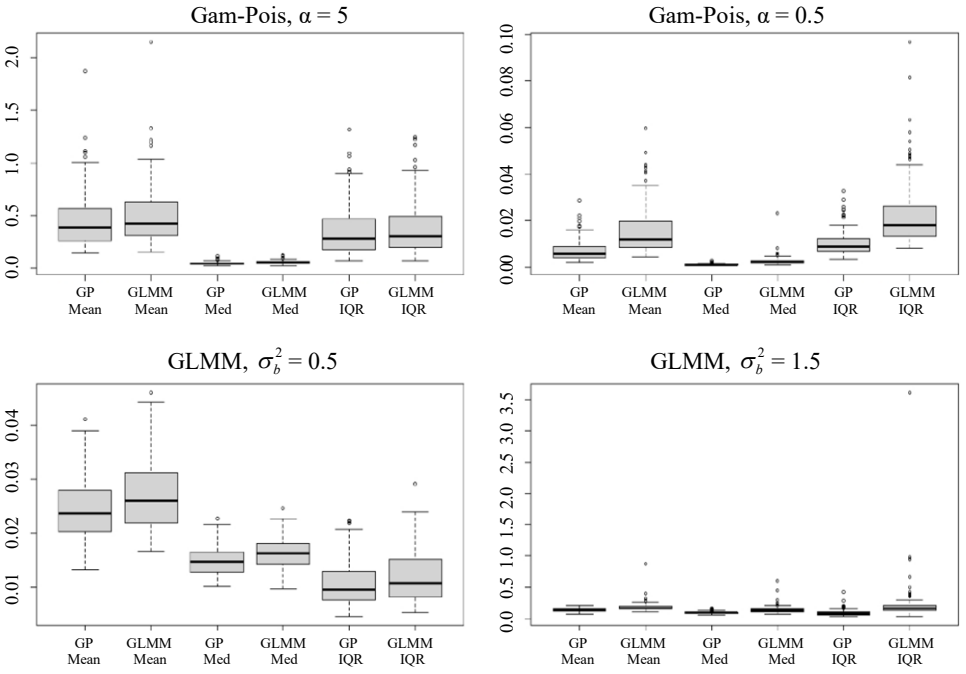
Figure 4.2 contains boxplots of the RRMSEMSE_i. The two MSE estimators have consistently similar relative root mean square errors. The biases observed for certain simulation configurations seem to have a negligible effect on the MSE of the MSE estimator.

Figure 4.1 Relative biases (RB_i) of MSE estimators for four simulation configurations.



Notes: GLMM = Generalized linear mixed model; GP = gamma-Poisson; IQR = Inter-quartile range; MSE = Mean square error.

Figure 4.2 Boxplots of relative root mean square errors of mean square error estimators (RRMSEMSE*i*) for four simulation configurations.



Notes: GLMM = Generalized linear mixed model; GP = gamma-Poisson; IQR = Inter-quartile range.

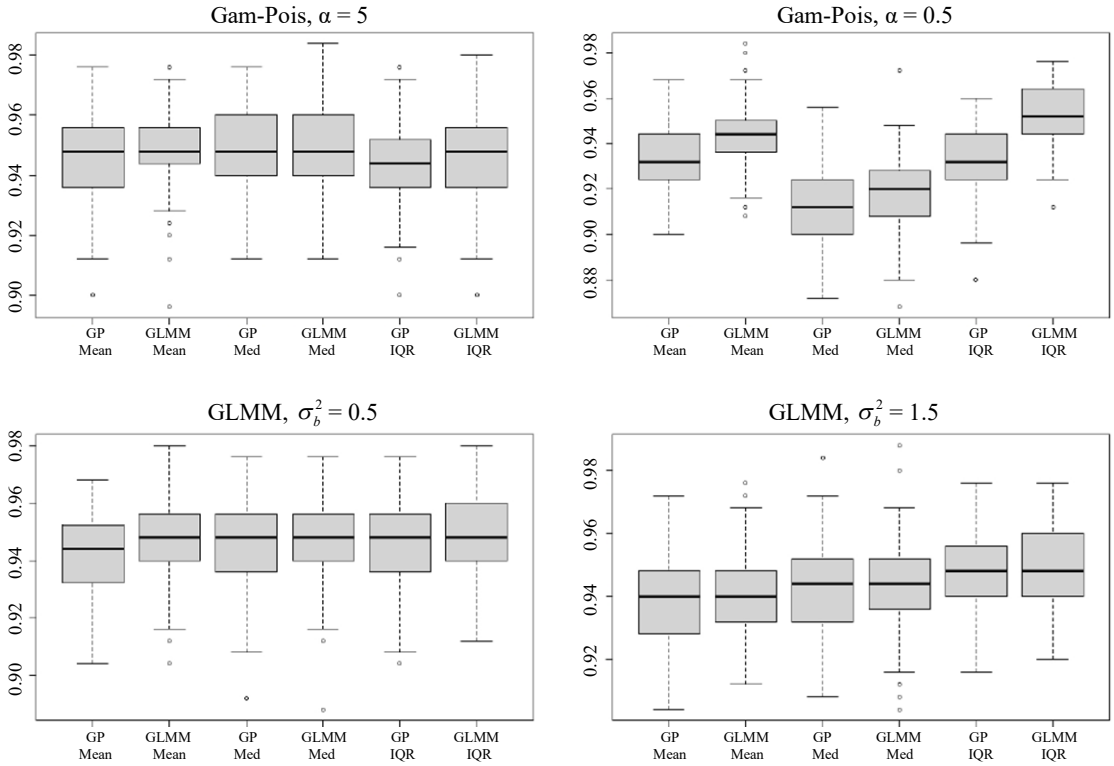
We define the empirical coverage of prediction intervals for area i by

$$\mathrm{CR}_i = \frac{1}{M}\sum_{m=1}^{M} I\left\{\theta_i^{(m)} \in \mathrm{CI}_i^{(m)}\right\},$$

where $\mathrm{CI}_i^{(m)} = \left[\hat{\theta}_i^{(m)} - 1.96\sqrt{\widehat{\mathrm{MSE}}_i^{(m)}}, \hat{\theta}_i^{(m)} + 1.96\sqrt{\widehat{\mathrm{MSE}}_i^{(m)}}\right]$. The empirical coverages of the nominal 95% prediction intervals are depicted in Figure 4.3. A surprising result is that the GLMM procedure tends to produce superior coverage rates than the Gam-Pois procedure under the Gam-Pois configuration with $\alpha =$ 0.5. Generally, the departures of the coverage rates from the nominal level are not severe. The empirical coverages of prediction intervals tend to fall between 92% and 98%.

In summary, the gamma-Poisson MSE estimator has reasonable properties when the gamma-Poisson model is the true model, and likewise, the GLMM MSE estimator has reasonable properties when the GLMM model is the true model. The properties of the MSE estimator under model misspecification depend on the parameter configuration. The GLMM MSE estimator is approximately unbiased under the Gam-Pois configuration when $\alpha = 5$ but has positive bias when $\alpha =$ 0.5. The Gam-Pois MSE estimator tends to have a median relative bias of about -5% under the GLMM configuration when $\sigma_b^2 =$ 0.5 and can be erratic when $\sigma_b^2 =$ 1.5.

Figure 4.3 Empirical coverages of nominal 95% prediction intervals (CR_i) for four simulation configurations.



Notes: GLMM = Generalized linear mixed model; GP = gamma-Poisson; IQR = Inter-quartile range.

5. Illustration with modeling observed vehicle occupants

We apply both the GLMM predictor and the Gam-Pois predictor to data from the 2018 Iowa Seat-Belt use survey. This is the same data set used in Berg (2022). While Berg (2022) only constructs predictors of means, we construct predictors of more general small area parameters.

The population consists of $N =$ 65,313 road segments in Iowa. The road segments are nested in $D = 15$ counties that define the small areas. The area sample sizes are $n_i = 5$ road segments for all but one county in which $n_i = 14$ such that the total sample size is $n = 84.$ Due to the small county sample sizes, this is clearly a small area estimation problem. Each road segment in the sample is observed for 45 minutes, and the response variable is defined as

y_{ij} = number of vehicle occupants observed during the 45-minute period on road segment j of county i.

Two covariates are available for every road segment in the population from the sampling frame. The first is the road type of the road segment, where the three road types are primary, secondary, and local. The second is the vehicle miles traveled. As in Berg (2022), we define the model to contain indicators for road type as well as interactions between road type and VMT. We refer to Berg (2022) for estimates of the fixed model parameters.

For this analysis, our objective is to compare the two models. In practice, however, an analyst may need to select one of the two models. We recommend diagnosing the goodness of fit of the two models using standardized residuals. We define a residual for the Gam-Pois model by

$$\frac{y_{ij} - \hat{\lambda}_{ij}\,\hat{u}_i}{\sqrt{\hat{\lambda}_{ij}\,\hat{u}_i}},$$

where

$$\hat{u}_i = \frac{y_{i.} + \hat{\alpha}}{\hat{\beta} + \hat{\lambda}_{i.}}.$$

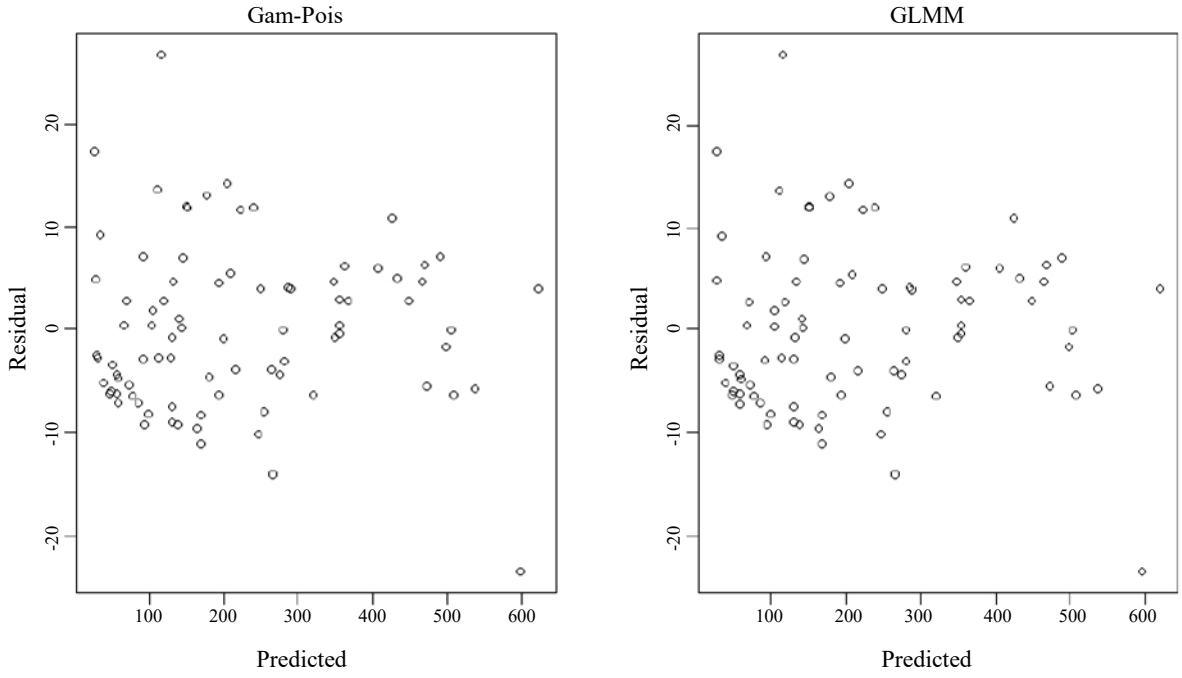
Note that $\hat{u}_i$ is an estimate of $E\left[u_i \,\middle|\, \mathbf{y}_{is}\right]$ (Berg, 2022). The standardized residual for the GLMM is defined as

$$\frac{y_{ij} - \hat{y}_{ij}^{\mathrm{PI}}}{\sqrt{\hat{y}_{ij}^{\mathrm{PI}}}}.$$

Figure 5.1 contains plots of the standardized residuals against the predicted values for the Gam-Pois and GLMM models. The residuals for the two models are strikingly similar. The standardized residuals do not exhibit systematic trends, and the variance of the residuals remains constant as the mean increases. The residuals clearly do not have a standard normal distribution; however, normality is not one of the model

assumptions. It is possible that the residuals indicate that the data are over-dispersed relative to the specified Poisson models. Incorporating over-dispersion in the proposed framework is a possible direction for future research.

Figure 5.1 Standardized residuals against predicted values for gamma-Poisson (gam-Pois) and generalized linear mixed models (GLMM).



We consider three county-level parameters of interest: the mean number of occupants per road segment, the median number of occupants, and the interquartile range of the number of occupants per road segment. We apply the methods of Sections 2 and 3 to obtain predictors and associated estimates of the mean square error. Because we are interested in parameters other than the mean, we use the predictor (2.4) for the gamma-Poisson model. We use the predictor (3.2) for the GLMM. We report the predictors and the coefficients of variation for each procedure. The coefficient of variation is the ratio of the square root of the estimated mean square error to the predictor. The MSE estimators are defined in (2.5) and (3.3), respectively, for the gamma-Poisson and GLMM models. We use a bootstrap sample size of $B = 200$. Table 5.1 contains the county level predictors and corresponding coefficients of variation based on the gamma-Poisson and GLMM models.

The two models produce consistently similar predictors. The predictors based on the gamma-Poisson model are nearly the same as the predictors based on the GLMM model. This empirical result is consistent with the finding in the simulation that both predictors tend to perform well, regardless of the true data generating model. For the data analysis, we do not know the "true" data generating model. Therefore, it is reassuring that the predictors based on the two models are similar.

Table 5.1
County predictors (pred) and coefficients of variation (cv) based on gamma-Poisson (Gam-Pois) and GLMM models.

	Mean				Median				IQR			
	Gam-Pois		GLMM		Gam-Pois		GLMM		Gam-Pois		GLMM	
County	pred	cv	pred	cv	pred	cv	pred	cv	pred	cv	pred	cv
1	60.37	0.46	59.99	0.49	34.71	1.38	34.27	1.46	104.8	0.26	103.56	0.28
2	60.29	0.04	60.18	0.04	26.31	0.05	26.54	0.10	57.86	0.14	58.40	0.14
3	62.97	0.1	63.01	0.09	24.54	0.11	24.53	0.08	63.58	0.55	63.70	0.56
4	87.82	0.45	87.60	0.45	46.34	1.25	46.31	1.26	163.53	0.3	163.52	0.31
5	54.67	0.24	54.91	0.23	24.5	1.05	24.39	1.03	60.26	0.76	59.94	0.76
6	77.08	0.17	77.47	0.17	31.23	0.99	31.50	0.97	136.85	0.07	137.57	0.08
7	65.79	0.15	66.03	0.13	30.27	0.34	30.20	0.30	67.68	0.63	67.32	0.61
8	99.65	0.02	99.34	0.03	41.53	0.05	41.56	0.05	118.65	0.21	118.05	0.23
9	68.12	0.17	68.19	0.16	25.98	0.2	26.07	0.17	73.11	0.36	73.14	0.37
10	97.24	0.27	97.29	0.27	41.26	0.42	41.40	0.40	153.54	0.36	154.04	0.38
11	96.99	0.06	96.83	0.05	41.9	0.09	42.02	0.07	91	0.18	91.07	0.18
12	55.73	0.05	56.14	0.05	24.39	0.3	24.47	0.26	97.05	0.1	97.15	0.10
13	53.43	0.04	53.09	0.05	18.36	0.07	18.19	0.13	53.52	0.22	53.12	0.23
14	122.55	0.16	123.69	0.15	47.28	0.17	47.78	0.15	131.63	0.38	132.73	0.39
15	96.98	0.09	96.89	0.09	48.34	0.52	48.28	0.53	91.33	0.41	91.40	0.42

Notes: GLMM = Generalized linear mixed model; IQR = Inter-quartile range.

The coefficients of variation based on the two models are also strikingly similar. This also reflects the results of the simulation in that the two procedures tend to produce reasonable means square error estimates, even under model misspecification. The coefficients of variation are not uniformly below 20%, a common threshold for determining an acceptable level of precision. Several of the coefficients of variation for the median and interquartile range exceed 30%. The coefficients of variation can exceed 100% for the median. For both the gamma-Poisson model and the GLMM, the effect of the variance due to estimating the fixed parameters on the mean square error of the small area predictors is substantial. This data set only has 15 counties. This leads to substantial variation in $\hat{\beta}_0$, $\hat{\sigma}_b^2$, $\hat{\alpha}$, and $\hat{\beta}$.

6. Discussion

We develop predictors of nonlinear parameters based on two unit level models for count data. We first define procedures for a gamma-Poisson model with unit-level covariates. We compare the gamma-Poisson model to a standard generalized linear mixed model. We use standard parametric bootstrap procedures for both models.

A limitation of the bootstrap procedure that we employ is that the bootstrap MSE estimator is not second-order unbiased. One can use the double bootstrap to construct a bias-corrected MSE estimator (Hall and Maiti, 2006a,b; Erciulescu and Fuller, 2014). We do not pursue the double bootstrap in this work because the proposed MSE estimator has adequate properties for the simulation configurations that we considered.

We study the empirical properties of the small area predictors when the model is correctly specified and under model misspecification. The main conclusion from the simulations is that the empirical best predictor

for the gamma-Poisson model is superior when the gamma-Poisson model is true, and the PI and GLMM predictors are superior when the GLMM is true, as expected. This illustrates the importance of validating the model assumptions in model-based small area estimation. However, an interesting result is that the loss of efficiency from using the GLMM or gamma-Poisson predictor when the other model is true is consistently small. The MSE estimators exhibit more bias when the model is incorrectly specified than when the model is correctly specified. Nonetheless, the relative biases of the MSE estimators are usually between -20% and 20%, regardless of which model is the true model. The coverage rates of confidence intervals do not exhibit severe and systematic over-coverage or under-coverage, even under model misspecification. The PI predictor has reasonable properties for the mean but is inefficient for nonlinear parameters.

The data analysis re-affirms the results of the simulation study. The gamma-Poisson and GLMM procedures lead to similar predictors in the data analysis. This result echoes an empirical finding of Clayton and Kaldor (1987) that estimates of lung cancer rates based on gamma-Poisson and Poisson-lognormal models are similar.

The gamma-Poisson model has two main strengths relative to the Poisson GLMM. An important benefit of the gamma-Poisson procedure is computational simplicity. Maximum likelihood estimation is easy for the gamma-Poisson model because calculation of the marginal likelihood does not require numerical integration. We do not use maximum likelihood estimation for the GLMM model. Instead, we use an IRLS algorithm (Schall, 1991) for computational simplicity. A formal comparison of maximum likelihood estimation to the IRLS algorithm for the purpose of small area estimation of count data is a possible future research direction. We present output using the `glmer R` function in Section 2 of the supplementary material (Berg, 2023b). Given estimates of the fixed parameters, the predictors are easier to calculate for the gamma-Poisson model than for the Poisson GLMM. The computing time to implement the gamma-Poisson predictor is roughly half the time required to implement the GLMM predictor (see Section 4 of the supplementary material (Berg, 2023b) for further detail). A second strength of the gamma-Poisson procedure is that the estimators of fixed model parameters are consistent. In contrast, the IRLS estimators for the Poisson GLMM are known to be inconsistent. The bootstrap procedure for the GLMM relies on consistency of the model parameter estimators, an assumption that does not hold for our estimation procedure. The procedures for the GLMM model require approximations, and we evaluate the validity of these approximations through simulation.

The GLMM exhibits different strengths, relative to the gamma-Poisson procedure. The loss of efficiency from incorrect use of the GLMM is slightly below the loss from incorrect use of the Gam-Pois model. The coverage rates of confidence intervals for the GLMM predictor are somewhat closer to 95% than the coverage rates for the Gam-Pois model for certain parameters and configurations.

In practice, the analyst may need to choose one of the two models. We propose to use residuals to diagnose the goodness of fit of the models. Our experience is that it can be very difficult to distinguish between the two models. The two models differ only with respect to the distribution of the random effect and generate very similar types of data. Fortunately, the analysis in this paper suggests that the proposed

methods are fairly robust to specification of the incorrect distributional form. The methods for the gamma-Poisson model work well when the Poisson GLMM is true and vice versa. Our analysis suggests that both the gamma-Poisson and Poisson GLMM will lead to similar results for many parameter configurations. We expect the conclusions for a specific application to be fairly insensitive to the choice of model. We encourage the analyst to consider the strengths and weaknesses of the two models for particular applications when selecting one of them in practice. We also note that an alternative to model selection is model averaging (Aitkin, Liu and Chadwick, 2009). An investigation of model averaging for the two models proposed here is a possible direction for future research.

Based on this analysis, we prefer the gamma-Poisson model for computational simplicity. The conjugate form of the model makes the predictor and MSE estimator straightforward to calculate. If one is only interested in the mean, then we recommend the closed-form predictor and MSE estimator of Berg (2022). For prediction of general parameters, we recommend the simulation-based predictor defined as $\hat{\theta}_{i,L}$ in (2.4). The gamma-Poisson predictor is straightforward to implement and has acceptable efficiency, even if the GLMM true.

Acknowledgements

This research was supported by USDA NRCS Cooperative Agreement, 68-7482-17-009.

References

Aitkin, M., Liu, C.C. and Chadwick, T. (2009). Bayesian model comparison and model averaging for small-area estimation. *Ann. Appl. Stat.*, 3(1), 199-221.

Berg, E. (2022). Empirical best prediction of small area means based on a unit-level gamma-poisson model. *Journal of Survey Statistics and Methodology*, 1-22.

Berg, E. (2023a). Small area prediction of seat-belt use rates using a bayesian hierarchical unit-level poisson model with multivariate random effects. *Stat*, 12(1), e544.

Berg, E. (2023b). Supplementary material for "Prediction of general small area parameters for unit-level count data". [Online]. Available from: https://github.com/emilyjb/Count-Data-Supplement.

Berg, E., and Eideh, A. (2023). Small area inference for exponential dispersion families under informative sampling. Manuscript under review by *Journal of Survey Statistics and Methodology*.

Boubeta, M., Lombardía, M.-J. and Morales, D. (2016). Empirical best prediction under area-level poisson mixed models. *Test*, 25(3), 548-569.

Clayton, D., and Kaldor, J. (1987). Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 671-681.

Erciulescu, A.L., and Fuller, W.A. (2014). Parametric bootstrap procedures for small area prediction variance. *Proceedings of the Survey Research Methods Section*, American Statistical Association Washington, DC, 3307-3318.

Ghosh, M. (2020). Small area estimation: Its evolution in five decades. *Statistics in Transition, New Series*, 21(4), 1-22.

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.

Guadarrama, M., Molina, I. and Rao, J.N.K. (2018). Small area estimation of general parameters under complex sampling designs. *Computational Statistics & Data Analysis*, 121, 20-40.

Hall, P., and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics*, 34, 4, 1733-1750.

Hall, P., and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2), 221-238.

Hobza, T., Marhuenda, Y. and Morales, D. (2020). Small area estimation of additive parameters under unit-level generalized linear mixed models. *SORT-Statistics and Operations Research Transactions*, 3-38.

Jiang, J., and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15(1), 1-96.

Marhuenda, Y., Molina, I., Morales, D. and Rao, J.N.K. (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.

Molina, I., Corral, P. and Nguyen, M. (2022). Estimation of poverty and inequality in small areas: Review and discussion. *TEST*, 1-24.

Molina, I., Nandram, B. and Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical bayes approach. *The Annals of Applied Statistics*, 8(2), 852-885.

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

Parker, P.A., Janicki, R. and Holan, S.H. (2019). Unit level modeling of survey data for small area estimation under informative sampling: A comprehensive overview with extensions. *arXiv preprint arXiv:1908.10488*.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1), 40-68.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Reluga, K., Lombardía, M.-J. and Sperlich, S. (2021). Simultaneous inference for empirical best predictors with a poverty study in small areas. *Journal of the American Statistical Association*, (accepted), 1-33.

Rojas-Perilla, N., Pannier, S., Schmid, T. and Tzavidis, N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(1), 121-148.

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

Smith, A.F., and Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46(2), 84-88.

Tzavidis, N., Ranalli, M.G., Salvati, N., Dreassi, E. and Chambers, R. (2015). Robust small area prediction for counts. *Statistical Methods in Medical Research*, 24(3), 373-395.