ORIGINAL RESEARCH

A Unit-Level One-Inflated Beta Model for Small Area Prediction of Seat-Belt Use Rates

Zirou Zhou[a] and Emily Berg[b]

[a, b]Department of Statistics, Iowa State University, 2438 Osborn Dr Ames, IA, USA

ARTICLE HISTORY
Compiled May 12, 2024

ABSTRACT
We develop a unit-level one-inflated beta model for the purpose of small area estimation. Our specific interest is in estimation of seat-belt use rates for Iowa counties using data from the Iowa Seat-Belt Use Survey. As a result of small county sample sizes, small area estimation methods are needed. We propose frequentist and Bayesian implementations of a unit-level one-inflated beta model. We compare the Bayesian and frequentist predictors to simpler alternatives through simulation. We apply the proposed Bayesian and frequentist procedures to data from the Iowa Seat-Belt Use Survey.

KEYWORDS
Small area estimation, Parametric Bootstrap, Bayesian, Beta Distribution, One-inflated

1. Introduction

The Iowa Seat-Belt Use Survey is an annual survey designed to estimate the state level seat-belt use rate. In this paper, we focus on estimating seat-belt use rates among drivers for the year 2018. The design of the seat-belt use survey is a two-stage design. At the first stage, a sample of counties is selected. At the second stage, a sample of road segments is selected from each sampled county. Each road segment is observed for 45 minutes. The proportion of drivers wearing a seat-belt is recorded for each sampled road segment. The road-segment level proportions are aggregated to produce an estimate of the seat-belt use rate for the state of Iowa. It is important to emphasize that the sampling units in the survey are road segments and not drivers. Therefore, the relevant units for analysis are the road segments. Incorrectly treating the drivers as the sampling units would lead to the potential to greatly understate measures of uncertainty.

Estimates of seat-belt use rates at the county level are of interest. As a result of small county sample sizes, county level estimates are not produced as part of the standard operation. Small area estimation methods are needed to produce the county estimates. Small area estimation uses model assumptions and auxiliary information to construct estimators for domains with small sample sizes. The use of valid statistical models can provide small area estimates with greater precision than the direct estimates. We

CONTACT Zirou Zhou Email: zzhou@iastate.edu

refer the reader to [12, 19, 21] for reviews of small area estimation.

Small area estimation requires defining an appropriate model for the observed data. In the Seat-Belt Survey application, the response variable is the proportion of belted drivers on a road segment. The proportions are in the interval $(0, 1]$. A one-inflated beta model is therefore a rational choice for the response distribution.

We develop a unit-level one-inflated beta model for small area estimation. The model incorporates unit-level covariates and area-level random effects. The Seat-Belt Survey motivates our interest in this model, but the model is applicable more broadly to any response variable with support $(0, 1]$. Further, the procedure extends immediately to data with support $[0, 1]$, as we explain in Section 7 of the Supplement. We focus on one-inflated data because this aligns with the support of the proportions observed in the seat-belt survey.

The road-segment level proportions are ratios of the number of belted drivers to the number of drivers observed on a road segment. In [2], we conduct an analysis of this data set that uses Poisson distributions for the raw, collected counts. A problem with the analysis of the counts is that the standard errors of the small area predictors are not uniformly below the standard errors of the direct estimators. The analysis of the proportions in this paper is an attempt to overcome the limitations of the analysis of the counts.

1.1. Literature Review

Our work relates to two overlapping areas of existing literature. The first concerns extensions of the Fay-Herriot area level model of [7] to a beta response distribution. The second concerns unit-level models for zero-inflated or one-inflated data.

Beta distributions with discrete components have been used in the context of area-level small area models. [11] studies the properties of an area-level small area model with a beta response distribution, and [22] extends the model to a multivariate area-level model. [24] use zero-one inflated beta distributions to model area-level estimates of poverty rates. [13] have used the area-level zero-inflated beta distribution to estimate poverty rates in villages in Indonesia. [6] uses area-level inflated beta models for small area estimation with local-global shrinkage priors for the random effects. While these works specify the model at the area level, we develop a unit-level model with a one-inflated beta response distribution.

One-inflation in the data makes the assumptions of a model for continuous random variables invalid ([17]), and therefore problems with inference are liable to occur by ignoring this feature of the data. In the classical regression literature, mixture models (also referred to as the two part models), which separately model the non-one values and the occurrence of one values, are widely used to account for excess ones in data. See [8, 15, 23]. Unit-level models for zero-inflated data have been developed in combination with normal, log normal, and gamma response distributions. [4, 20] develop small area estimation procedures for unit-level zero-inflated data, where the continuous component of the distribution is a normal distribution. It is noteworthy that [20] used a full two-part random effects model that accounted for the correlations between random effects of the two parts of the model (i.e., LMM and GLMM). However, their simulation results showed that this correlation does not substantively improve the small area predictors. Moreover, use of this correlation makes model fitting computationally intensive and not always stable. We will not incorporate this type of correlation in our modeling, but we will evaluate the sensitivity of our pre-

dictors to the presence of this correlation through simulation. For skewed, positive response variables, [16] develops a zero-inflated log normal model, and [5] develops a zero-inflated gamma model. We extend the small area literature for semi-continuous unit-level data to a one-inflated unit-level beta regression model.

1.2. Overview of Manuscript

We develop a unit-level one-inflated beta regression model for the purpose of small area estimation. In Section 2, we motivate the model through an exploratory data analysis. In Section 3, we define the small area model and provide both Bayesian and frequentist inference procedures. In Section 4, we define the two main competing predictors. One is the empirical best linear unbiased predictor (EBLUP) for the linear unit-level model of [1]. The other is a one-inflated version of the model of [4]. In Section 5, we evaluate the properties of the procedure through simulation and compare the proposed predictor to the competitors. We apply the unit-level one-inflated beta model to data from the Iowa Seat-Belt Use Survey in Section 6. We discuss the strengths and weaknesses of the Bayesian and frequentist approaches in Section 7.

2. Exploratory Analysis of the 2018 seat-belt data

In this section, we examine the proportions of belted drivers on the sampled road segments. We also examine the relationships between these proportions and the possible model covariates. One covariate is the annual average vehicle miles traveled across the road segment in a given year, abbreviated VMT. The other covariate is the road type, where the three road types are primary, secondary, and local roads. The road type and VMT are obtained from administrative data and are known for every road segment in the population from the sampling frame. In preparation for defining a model, we explore the nature of the available data in this section.

2.1. Histograms of Proportions

Figure 1 shows a histogram of the proportions of belted drivers on the sampled road segments. The distribution of proportions is very left skewed, and many of the proportions are close or equal to 1. Thus, use of a unit-level one-inflated beta model appears reasonable for this data set.

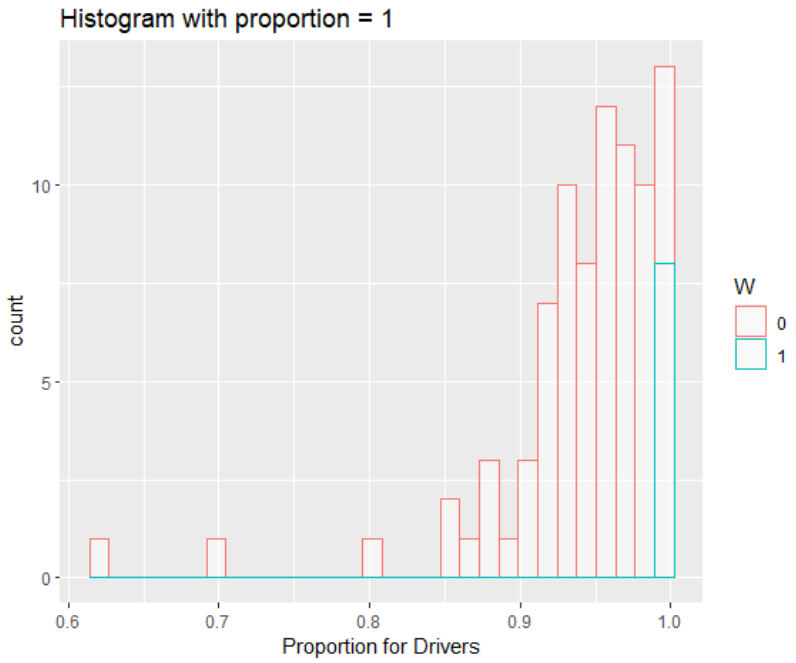

Figure 1. Histograms of Proportions for all observations, where W identifies whether a proportion is 1

2.2. *Scatterplot of Proportions against log VMT*

Figure 2 contains a scatterplot of the logits of the proportions against the log VMT. A small constant (0.01) is subtracted from the proportions so that the logit transformation is defined for proportions that are equal to one. The red line is the basic linear regression of the logits of the proportions against log VMT, where we first subtract the small constant from the proportions before defining the logits. As the scatterplot shows, the log VMT and logits of the proportions have a strong, linear association.

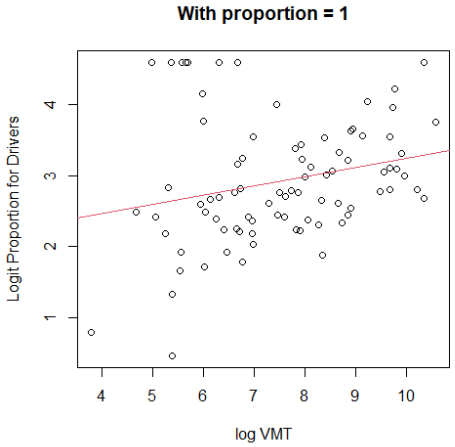

Figure 2. Scatter plot of empirical logits of proportions against log VMT

2.3. *Box plot of Proportions against Road.Type*

Figure 3 contains a boxplot of the proportions for the three road types. As the boxplot shows, the proportions vary for the different road types. Not only the means, but also the standard deviations, vary between the groups.

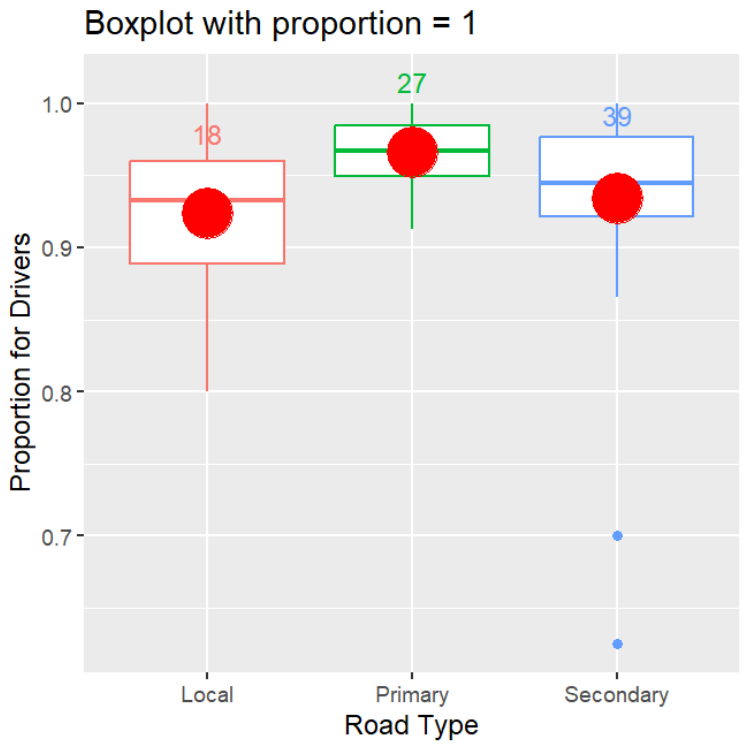

Figure 3. Boxplot of Proportions against Road Type

2.4. *Implications of Exploratory Analysis*

Based on the exploratory analysis, the unit-level beta distribution with one-inflation appears to be a natural model to use for this data set. This distribution preserves the support of the observed proportions. The beta model also allows the variance to change with the mean, thereby reflecting the nonconstant variances observed in Figure 3. In the next section, we define frequentist and Bayesian implementations of the one-inflated beta model.

3. Model and Predictor

Let $i = 1, \ldots, D$ index the areas and $j = 1, \ldots, N_i$ index the units in the population for area i. Let y_{ij} be the variable of interest for unit j in area i. The support of y_{ij} is $(0, 1]$. In the seat-belt survey application, y_{ij} represents the proportion of belted drivers on road segment j of the county i. Although multiple drivers are observed on a road segment, the road segments are units in the sense that they are sampling units nested in counties.

Assume y_{ij} satisfies a one-inflated beta regression model. To define the model, let

$Beta(\mu_{ij}, \phi)$ denote a random variable having a Beta distribution with shape parameters $\mu_{ij}\phi$ and $(1-\mu_{ij})\phi$. Then, assume

$$y_{ij} = \begin{cases} 1 & \text{with prob. } p_{ij} \\ Beta(\mu_{ij}, \phi) & \text{with prob. } (1-p_{ij}), \end{cases} \tag{1}$$

where

$$\text{logit}(p_{ij}) = \boldsymbol{z}'_{ij}\boldsymbol{\alpha} + u_i, \tag{2}$$

$$\text{logit}(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + b_i \tag{3}$$

$u_i \overset{iid}{\sim} N(0, \sigma_u^2)$, $b_i \overset{iid}{\sim} N(0, \sigma_b^2)$, and $Cov(u_i, b_i) = 0$. The parameter of interest is the area mean defined by

$$\bar{y}_{N_i} = \frac{\sum_{j \in U_i} \omega_{ij} y_{ij}}{\sum_{j \in U_i} \omega_{ij}}, \tag{4}$$

where U_i is the index set of the N_i elements for area i. The ω_{ij} is a weight that is known for every element of the population. In the seat-belt survey application, ω_{ij} is the vehicle miles traveled for road segment j of county i. The vehicle miles traveled is known for every road segment in the population from the sampling frame. In the common situation, $\omega_{ij} = 1$.

We do not observe y_{ij} for the full population. We only observe y_{ij} for a sample A_i of n_i elements for area i. We let $j = 1, \ldots, n_i$ index the elements in the sample for area i, such that $j = n_i + 1, \ldots, N_i$ indexes the nonsampled elements. We assume that the covariates $(\boldsymbol{z}_{ij}, \boldsymbol{x}_{ij})$ are available for all elements of the population. This is a common requirement of nonlinear small area methods, and the assumption that the covariates are known for the full population is satisfied for the Seat-Belt Use Survey. We develop frequentist and Bayesian inference procedures.

We focus on the model for one-inflated data because the proportions in the Seat-Belt Use Survey are in the interval $(0, 1]$. The model generalizes easily to data with support $[0, 1]$. We extend the model to zero-one-inflated data in Section 6 of the Supplement.

3.1. Frequentist Inference Procedure

We define the steps of a frequentist procedure for constructing a predictor of $\bar{y}_{N_i}$ from the available data.

(1) Maximize a Laplace approximation to the likelihood corresponding to the beta distribution to obtain an estimate of $(\boldsymbol{\beta}', \phi)'$, and maximize the complete data likelihood, evaluated at the estimates from the Laplace approximation, to obtain the predictor of b_i. We explain the estimation procedure in more detail in Section 6 of the Supplement. Operationally, we use the `R` function `glmmTMB` to obtain an estimator $(\hat{\boldsymbol{\beta}}', \hat{\phi})'$ and a predictor $\hat{b}_i$. [3, 14] provide further information on the `glmmTMB` estimation procedure.

(2) Maximize a Laplace approximation to the likelihood for the binary component to obtain an estimate $\hat{\boldsymbol{\alpha}}$ and a predictor $\hat{u}_i$. We implement estimation for the

binary part using the `R` function `glmer`.

(3) For every $i = 1, \ldots, D$ and $j \in U_i$, define a predictor of μ_{ij} by

$$\hat{\mu}_{ij} = \frac{\exp(\boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}} + \hat{b}_i)}{1 + \exp(\boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}} + \hat{b}_i)}. \tag{5}$$

(4) For every $i = 1, \ldots, D$ and $j \in U_i$, define a predictor of p_{ij} by

$$\hat{p}_{ij} = \frac{\exp(\boldsymbol{z}'_{ij}\hat{\boldsymbol{\alpha}} + \hat{u}_i)}{1 + \exp(\boldsymbol{z}'_{ij}\hat{\boldsymbol{\alpha}} + \hat{u}_i)}. \tag{6}$$

(5) For $i = 1, \ldots, D$ and $j \in U_i$, define a predictor of y_{ij} by

$$\hat{y}_{ij} = (1 - \hat{p}_{ij})\hat{\mu}_{ij} + \hat{p}_{ij}. \tag{7}$$

(6) Define a predictor of $\bar{y}_{N_i}$ by

$$\hat{\bar{y}}_{N_i} = \frac{1}{\sum_{j \in U_i} \omega_{ij}} \sum_{j \in U_i} \omega_{ij} \hat{y}_{ij}. \tag{8}$$

In the common situation in which $\omega_{ij} = 1$, the predictor is of the form $\hat{\bar{y}}_{N_i} = N_i^{-1} \sum_{j \in U_i} \hat{y}_{ij}$.

The predictor (8) implicitly assumes that the finite population correction factor is negligible. If the sampling fraction is important, then one can define a predictor as

$$\hat{\bar{y}}_{N_i} = \frac{1}{\sum_{j \in U_i} \omega_{ij}} \left\{ \sum_{j=1}^{n_i} \omega_{ij} y_{ij} + \sum_{j=n_i+1}^{N_i} \omega_{ij} \hat{y}_{ij} \right\}.$$

In the seat-belt survey, the sampling fractions are small, so we focus on the predictor (8).

3.1.1. Bootstrap MSE Estimation

We use the bootstrap to estimate the mean square error (MSE) of the proposed frequentist predictor (8). The bootstrap procedure is an application of the parametric bootstrap method of [10] to the unit-level one-inflated beta model. We define the bootstrap MSE estimator in this section. For $b = 1, \ldots, B$, we repeat the following steps:

(1) Simulate a population from the model defined by (1)-(3) with parameters equal to the estimated parameters, $\hat{\boldsymbol{\beta}}$, $\hat{\phi}$, $\hat{\boldsymbol{\alpha}}$, $\hat{\sigma}_b^2$, and $\hat{\sigma}_u^2$. For $i = 1, \ldots, D$, let $y_{i1}^{(b)}, \ldots, y_{iN_i}^{(b)}$ denote the bootstrap population. Define the bootstrap version of the population parameter by $\bar{y}_{N_i}^{(b)} = (\sum_{j \in U_i} \omega_{ij})^{-1} \sum_{j \in U_i} \omega_{ij} y_{ij}^{(b)}$. Let $y_{i1}^{(b)}, \ldots, y_{in_i}^{(b)}$ denote the simulated response variables corresponding to sampled elements. Call $\{y_{ij}^{(b)} : j = 1, \ldots, n_i\}$ the bootstrap sample.

(2) Using the bootstrap sample, repeat the proposed prediction procedure. Let $\hat{\bar{y}}_{N_i}^{(b)}$ be the predictor of (8) constructed with the bootstrap sample.

Define the MSE estimator for area i by

$$\widehat{MSE}_i = B^{-1} \sum_{b=1}^{B} (\hat{\bar{y}}_{N_i}^{(b)} - \bar{y}_{N_i}^{(b)})^2.$$

We construct a normal theory 95% prediction interval as $\hat{\bar{y}}_{N_i} \pm 1.96\sqrt{\widehat{MSE}_i}$.

3.2. Bayesian Inference Procedures

We next define a Bayesian predictor of $\bar{y}_{N_i}$. As in the model from (1)-(3), the parameter vector $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \sigma_u^2, \sigma_b^2, \phi)$ needs to be estimated, and we need to predict the random variables $(u_i, b_i)'$, which are realizations from normal distributions. To complete the Bayesian specification, we require prior distributions for $\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_u^2, \sigma_b^2$, and ϕ. We define the priors as $\boldsymbol{\alpha} \sim N(\mathbf{0}, 10^4 \boldsymbol{I}_{p\times p})$, $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^4 \boldsymbol{I}_{q\times q})$, $\phi \sim$ Half-Cauchy$(0, 10000)$, $\sigma_b^2 \sim$ Half-Cauchy$(0, 10000)$, and $\sigma_u^2 \sim$ Half-Cauchy$(0, 10000)$, where p is the dimension of $\boldsymbol{\beta}$, q is the dimension of $\boldsymbol{\alpha}$, and the notation Half-Cauchy$(0, 10000)$ denotes a Half-Cauchy distribution with a scale parameter of 10000. The reference [9] recommends the half-Cauchy prior for positive-valued parameters, as it is part of the conditionally-conjugate, folded-noncentral-t distribution family. We opt for a large scale parameter in this prior because our prior information is limited, and a large scale parameter leads to a more diffuse prior. All priors are assumed to be mutually independent. The posterior is then of the form

$$\begin{aligned}
&\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_b^2, \sigma_u^2, \phi, u_i, b_i \mid \boldsymbol{y}, \boldsymbol{w}) \\
&\quad \propto [\prod_{i=1}^{D} \{\prod_{j=1}^{n_i} f_{ij}(y_{ij}, w_{ij} \mid b_i, u_i, \phi)\} \sigma_b^{-1} \phi(b_i/\sigma_b) \sigma_u^{-1} \phi(u_i/\sigma_u)] \pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \phi, \sigma_b^2, \sigma_u^2)
\end{aligned}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_D')'$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$ $\boldsymbol{w} = (\boldsymbol{w}_1', \ldots, \boldsymbol{w}_D')'$, $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{in_i})'$,

$$w_{ij} = \begin{cases} 1 & \text{if } y_{ij} = 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$f_{ij}(y_{ij}, w_{ij} \mid b_i, u_i, \phi) = \{f_{ij}(y_{ij} \mid b_i, \phi)\}^{(1-w_{ij})} p_{ij}^{w_{ij}} (1 - p_{ij})^{1-w_{ij}},$$

$f_{ij}(y_{ij} \mid b_i, \phi)$ is the density of a beta distribution with shape parameters $\phi\mu_{ij}$ and $\phi(1 - \mu_{ij})$, and $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \phi, \sigma_b^2, \sigma_u^2)$ is the density corresponding to the specified prior distributions. We use `RStan` and `INLA` to sample from the posterior distribution. This results in samples $\boldsymbol{\alpha}^{(\ell)}$, $\boldsymbol{\beta}^{(\ell)}$, $\phi^{(\ell)}$, $u_i^{(\ell)}$, and $b_i^{(\ell)}$ for $\ell = 1, \ldots, L$, where ℓ denotes the simulated sample from the posterior distribution.

We then use the method of [18] to construct the predictors and the uncertainty measures. [18] is a very general procedure for conducting Bayesian inference in the

context of small area estimation. For each simulated sample, ℓ, from the posterior distribution, we generate for $j = 1, \ldots, N_i$,

$$\begin{aligned} y_{ij}^{*(\ell)} &\sim \text{Beta}(\mu_{ij}^{(\ell)}, \phi^{(\ell)}), \\ w_{ij}^{(\ell)} &\sim \text{Bernoulli}(p_{ij}^{(\ell)}), \\ \text{logit}(\mu_{ij}^{(\ell)}) &= \boldsymbol{x}_{ij}'\boldsymbol{\beta}^{(\ell)} + b_i^{(\ell)}, \\ \text{logit}(p_{ij}^{(\ell)}) &= \boldsymbol{z}_{ij}'\boldsymbol{\alpha}^{(\ell)} + u_i^{(\ell)}. \end{aligned}$$

We then set $y_{ij}^{(\ell)} = w_{ij}^{(\ell)} + (1 - w_{ij}^{(\ell)})y_{ij}^{*(\ell)}$. We then define a sample from the posterior predictive distribution of $\bar{y}_{N_i}$ given the observed data as

$$\bar{y}_{N_i}^{(\ell)} = \frac{\sum_{j \in U_i} \omega_{ij} y_{ij}^{(\ell)}}{\sum_{j \in U_i} \omega_{ij}}.$$

This results in simulated samples $\bar{y}_{N_i}^{(\ell)}$ for $\ell = 1, \ldots, L$. We define the Bayesian predictor as the posterior mean given by

$$\hat{\bar{y}}_{N_i}^{B} = \frac{1}{L} \sum_{\ell=1}^{L} \bar{y}_{N_i}^{(\ell)}. \tag{9}$$

To assess the uncertainty in the predictor, we use the posterior mean square error (MSE) defined as

$$\widehat{MSE}_i^{B} = \frac{1}{L-1} \sum_{\ell=1}^{L} (\bar{y}_{N_i}^{(\ell)} - \hat{\bar{y}}_{N_i}^{B})^2. \tag{10}$$

As (10) is the sample variance of $\{\bar{y}_{N_i}^{(\ell)} : \ell = 1, \ldots, L\}$, we refer to $\sqrt{\widehat{MSE}_i^{B}}$ as a posterior root MSE and as a posterior standard deviation interchangeably. We construct a normal theory confidence interval as $\hat{\bar{y}}_{N_i}^{B} \pm 1.96\sqrt{\widehat{MSE}_i^{B}}$. Similar to the frequentist predictor, the Bayesian predictor assumes that the finite population correction factor is negligible. If the sampling fraction is important, one can set $y_{ij}^{(\ell)} = y_{ij}$ for $j = 1, \ldots, n_i$.

3.2.1. Bayesian Computation Procedures

We compare two methods of computing the Bayesian predictors. The first method is Stan and the second is Integrated Nested Laplace Approximation (INLA). We use the same priors for both procedures. Stan is a traditional MCMC algorithm that primarily uses the No-U-Turn Sampler (NUTS), an advanced form of Hamiltonian Monte Carlo (HMC). These samplers effectively navigate the parameter space of the model to generate samples from the posterior distribution. INLA is a sophisticated method for conducting approximate Bayesian inference, particularly well-suited to models that can be expressed as latent Gaussian Markov random fields (GMRF). At

the heart of INLA is the Laplace approximation technique used to estimate complex integrals quickly and effectively. By using Taylor series expansions around the mode of functions, INLA can approximate the log of the posterior density, simplifying the calculation of posterior marginals. One of the main advantages is the use of GMRFs and deterministic numerical integration allows INLA to perform Bayesian inference much faster, making it a practical choice for complex models that need quick inferences.The algorithm first approximates the marginal posterior distributions of the hyperparameters. It then uses these approximations to calculate the marginal posteriors of the latent variables by integrating over the hyperparameters' distributions.

4. Alternative methods

We consider several alternatives to the unit-level one-inflated beta distribution. The first alternative is the standard predictor of the small area mean for a unit-level linear model. The second extends the unit-level linear model to have a one-inflated component. The third takes a different approach and models the counts of belted drivers as realizations from binomial distributions conditional on the observed total number of drivers. We describe these alternative predictors in Sections 4.1-4.3.

4.1. The most common method: EBLUP

The first alternative is the empirical best linear unbiased predictor (EBLUP) for the unit-level linear model. The unit-level linear mixed model is first proposed in [1] for predicting county crop areas. If one ignores the issue that the support of the data is $(0, 1]$, then the unit-level linear mixed model would be the simplest small area predictor to use. We therefore compare the predictors proposed in Section 3 to the EBLUP for the unit-level linear mixed model. Chapter 7 of [21] provides a detailed account of small area prediction under the unit-level linear mixed model. We implement the EBLUP using the function `eblupBFH` in the `SAE R` package. When using this package, we estimate the variance components using REML.

4.2. Gaussian One-Inflated Model

The EBLUP of Section 4 is naive in the sense that it ignores the one-inflation in the data and the support of the continuous component. We consider a predictor that provides a compromise between the naive EBLUP and the proposed predictor based on the one-inflated beta model. We consider a one-inflated model with a Gaussian continuous component. The model is a slight modification of the procedure of [4] to handle one-inflated data instead of zero-inflated data.

Express the observed response variable as,

$$y_{ij} = z_{ij}(1 - \delta_{ij}) + \delta_{ij},$$

where z_{ij} satisfies the linear mixed effects model of Section 4.1 and δ_{ij} satisfies the logistic mixed model of Section 2. The predictor is of the form

$$\hat{y}_{ij}^{(3)} = \hat{z}_{ij}(1 - \hat{p}_{ij}) + \hat{p}_{ij}, \tag{11}$$

where $\hat{p}_{ij}$ is the predicted probability that δ_{ij} is 1 and $\hat{z}_{ij}$ is the EBLUP of z_{ij} based on the linear unit-level model for the observed values that do not equal one. The predictors $\hat{z}_{ij}$ and $\hat{p}_{ij}$ are constructed as in [4]. We then define the predictor as

$$\hat{\bar{y}}_{N_i}^{(3)} = \frac{1}{\sum_{j \in U_i} w_{ij}} \sum_{j=1}^{N_i} w_{ij} \hat{y}_{ij}^{(3)}, \tag{12}$$

where $\hat{y}_{ij}^{(3)}$ is defined in (11).

4.3. Binomial model

A different approach is to model the observed counts of belted drivers, conditional on the observed values for the total number of drivers. For this approach, we use a Bayesian binomial model. To define the model, let m_{ij} denote the number of belted drivers observed on road segment j of county i, and let M_{ij} be the corresponding total number of drivers. Let $Binomial(p_{ij}, M_{ij})$ denote a binomial distribution with sucess probability p_{ij} and a fixed binomial sample size of M_{ij}. Then, assume that

$$m_{ij} \overset{ind}{\sim} Binomial(p_{ij}, M_{ij}),$$

where $\text{logit}(p_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\alpha} + u_i$, and $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$.

A limitation of this model is that the M_{ij} are regarded as fixed. In the seat-belt survey application, the number of observed drivers is a random variable. The binomial model incorrectly treats these observed counts as fixed. By improperly regarding the binomial sample size as fixed, we run the risk of understating the uncertainty associated with the predictors.

Despite this limitation, we tried applying the binomial model to the data set in Section 8 of the Supplement. We find evidence of systematic bias in the predictors. Furthermore, the mean square errors are unreasonably high. Because this model performed poorly for our data set, we do not consider this model further.

5. Simulation Set-up

We use a parameter configuration for the simulation that is based loosely on the data for the Iowa Seat-Belt Use Survey. We set the number of areas to $D = 20$. For each small area, the population size is $N_i = 100$. We select a sample of $n_i = 10$ elements from the population of $N_i = 100$ elements in area i. We use the sample to estimate the parameters and construct predictors.

We generate y_{ij} as in (1). According to the formulas (2) and (3), we generate $x_{ij} \sim N(7.5, 1.5)$ and $z_{ij} \sim N(5, 2)$. The values for (α_0, α_1), (β_0, β_1) and ϕ will be set as $(2.16, -0.7162)$, $(1.25, 0.19)$ and 42, respectively. Also, the values for σ_b and σ_u will be set as 0.0957 and 0.277, respectively. The correlation between u_i and b_i will be set as 0 or 0.2, respectively. These parameters are based on a preliminary analysis of the real survey data, where we fit a model using only the vehicle miles traveled as the covariate.

5.1. *Comparison of Predictors*

We compare the predictors based on the zero-inflated beta model (proposed in Section 3) to the alternatives defined in Section 4.1 and 4.2. (We do not compare to the binomial model because this model performed poorly for our data set.) We evaluate the predictors on the basis of the Monte Carlo (MC) MSE. Let m denote the simulation iteration, where $m = 1, \ldots, M$. Let $\hat{\bar{y}}_{N_i}^{(1,m)}$ denote the direct estimates from the 10 samples observed. Let $\hat{\bar{y}}_{N_i}^{(2,m)}$ denote the frequentist predictor (8) based on the one-inflated beta model in MC simulation m. Let $\hat{\bar{y}}_{N_i}^{(3,m)}$ denote the EBLUP for the unit-level linear mixed model. When implementing the EBLUP, we include both $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ in the covariate vector. Let $\hat{\bar{y}}_{N_i}^{(4,m)}$ denote the predictor (12) for the one-inflated linear mixed model. Let $\hat{\bar{y}}_{N_i}^{(5,m)}$ denote the Stan-Bayesian predictor (9) for the unit-level one-inflated beta model. Let $\hat{\bar{y}}_{N_i}^{(6,m)}$ denote the INLA-Bayesian predictor (9) for the unit-level one-inflated beta model. Let $\bar{y}_i^{(m)}$ be the true area mean generated in MC simulation m. We define the average RMSE of predictor k for $k = 1, 2, 3, 4, 5, 6$ as

$$\text{RMSE}_k = \sqrt{\frac{1}{D}\sum_{i=1}^{D} \text{MSE}_{k,i}} = \sqrt{\frac{1}{D}\sum_{i=1}^{D}\frac{1}{M}\sum_{m=1}^{M}(\hat{\bar{y}}_{N_i}^{(k,m)} - \bar{y}_i{}^{(m)})^2}. \tag{13}$$

Observe that the MSE is defined as an average across areas and MC simulations. We use an MC sample size of $M = 5000$. We further define the improvement ratio as $\frac{RMSE_k}{RMSE_1}$ for $k \in \{1, 2, 3, 4, 5, 6\}$. The improvement ratio measures the increase (or decrease) in the MSE of an alternative predictor, relative to the frequentist predictor based on the one-inflated beta model.

5.2. *Simulation results*

Table 1 contains the average MSE as well as the improvement ratios for the alternative predictors when the correlation between u_i and b_i is zero. The alternatives defined in Section 3 are less efficient than the predictors based on the unit-level one-inflated beta distribution. The increase in MSE from using the simple EBLUP instead of the proposed frequentist predictor is 27%. The loss of efficiency from incorrect use of the Gaussian distribution in the one-inflated framework is 5%. The extra effort to model the data as having a one-inflated beta distribution appears worthwhile when the support of the observations is $(0, 1]$. The proposed frequentist predictor is also more efficient than the direct estimator, in terms of having a smaller average mean square error. This exemplifies the gain in efficiency from small area modeling that is widely documented in the small area literature. Model-based estimators are more efficient than direct estimators because they incorporate covariates and stronger modeling assumptions. One of the first illustrations of the benefits of modeling, relative to direct estimators, is [1]. Challenges arose when implementing the Bayesian model in Stan for the simulation study. Figures 4 and 5 contain histograms of the average deviations between the predicted values and the true means for the Stan and INLA procedures. (The average deviations are defined as $\sqrt{\frac{1}{D}\sum_{i=1}^{D}(\hat{\bar{y}}_i^{(k,m)} - \bar{y}_i{}^{(m)})^2}$ for $k = 5$ (Stan) and $k = 6$ (INLA), where $m = 1, \ldots, M$.) As illustrated in Figure 4, a few

samples from Stan had very extreme values for the predictors. These extreme values cause the MSE of the Bayesian predictor based on Stan to exceed the MSE of the frequentist predictor. If we remove these extreme values, then the MSE of the Stan-Bayesian predictor is slightly below that of the frequentist predictor. Use of the INLA procedure rectifies the problems associated with the Stan computational approach. When using INLA, we obtained no extreme samples. This can be seen by comparing the average deviations for Stan in Figure 4 to the corresponding average deviations for INLA in Figure 5. When the correlation is zero, the INLA-Bayesian predictor is slightly more efficient than the frequentist model-based predictor.

Table 2 presents the results for the configuration in which the correlation between u_i and b_i is 0.2. The results for this configuration are largely similar to the results with zero correlation. The main difference is that the loss of efficiency from the linear EBLUP is negligible when u_i and b_i are correlated. However, the naive EBLUP may produce predictors that are outside the parameter space for seat-belt use rates. Therefore, the linear EBLUP is not a viable alternative for this application. Another slight difference is that the INLA-Bayesian predictor suffers a slight loss of efficiency (of about 10%) relative to the frequentist predictor for the one-inflated beta model when the correlation is 0.2.

Table 1. Average RMSE and improvement ratio of alternative predictors for data with correlation = 0.0

Model Method	RMSE×100	Improvement Ratio
1. Direct Estimation	1.462	1.525
2. Frequentist: Unit-level One-inflated Beta	0.959	1
3. Linear EBLUP	1.086	1.132
4. One-inflated linear-mixed model	0.985	1.027
5. Stan-Bayesian: Unit-level One-inflated Beta	1.694	1.766
5*. Stan-Bayesian: Unit-level One-inflated Beta (remove extreme values)	0.954	0.995
6. INLA-Bayesian Unit-level One-inflated Beta	0.954	0.995

Table 2. Average RMSE and improvement ratio of alternative predictors for data with correlation = 0.2

Model Method	RMSE×100	Improvement Ratio
1. Direct Estimation	1.435	1.565
2. Frequentist: Unit-level One-inflated Beta	0.917	1
3. Linear EBLUP	0.922	1.005
4. One-inflated linear-mixed model	0.938	1.023
5. Stan-Bayesian: Unit-level One-inflated Beta	1.606	1.751
5*. Stan-Bayesian: Unit-level One-inflated Beta (remove extreme values)	0.860	0.937
6. INLA-Bayesian Unit-level One-inflated Beta	0.970	1.058

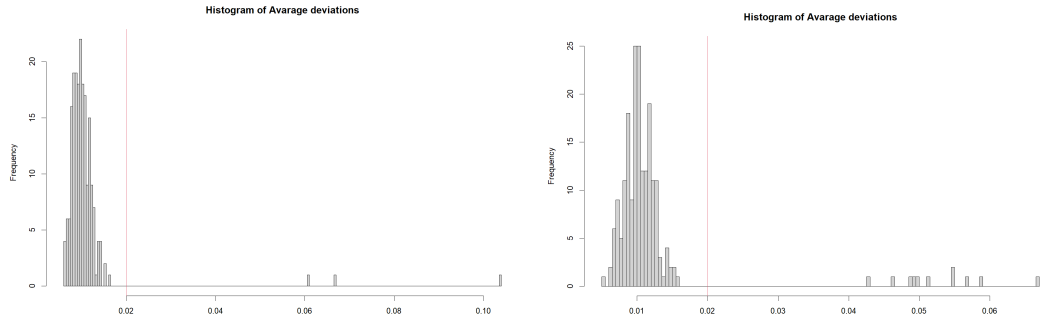

Figure 4. Histogram of the average deviations for Stan, defined by $\sqrt{\frac{1}{D}\sum_{i=1}^{D}(\hat{\bar{y}}_i^{(5,m)} - \bar{y}_i^{(m)})^2}$ for $i = 1, \ldots, M$. Value above 0.02 are considered extreme. Left panel is correlation = 0, and right panel is correlation = 0.2

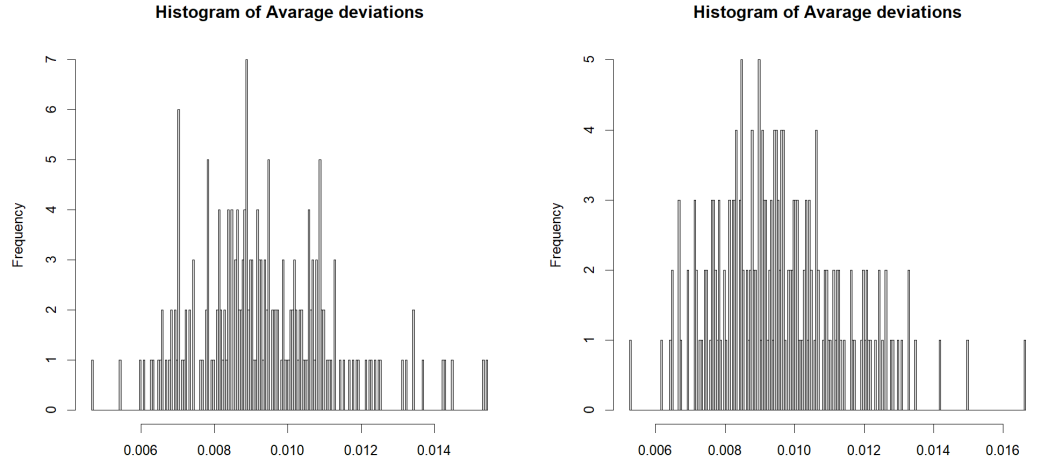

Figure 5. Histogram of the average deviations for INLA, defined by $\sqrt{\frac{1}{D}\sum_{i=1}^{D}(\hat{\bar{y}}_i^{(6,m)} - \bar{y}_i^{(m)})^2}$ for $i = 1, \ldots, M$. Left panel is correlation = 0, and right panel is correlation = 0.2

5.3. Bootstrap for Unit-level One-inflated Beta Model

As the frequentist predictor for the unit-level one-inflated beta model is among the most efficient predictors, we will next assess the bootstrap estimate of the MSE of this predictor. We use the bootstrap procedure described in Section 3.1. We use a bootstrap sample size of $B = 200$. We define the average relative bias of the bootstrap MSE estimator by

$$RB = \frac{MSE_B - MSE_1}{MSE_1},$$

where $\text{MSE}_1 = \frac{1}{D}\sum_{i=1}^{D}\frac{1}{M}\sum_{m=1}^{M}(\hat{\bar{y}}_{N_i}^{(1,m)} - \bar{y}_i^{(m)})^2$,
$MSE_B = M^{-1}\sum_{m=1}^{M} D^{-1}\sum_{i=1}^{D}\widehat{MSE}_i^{(m)}$, and $\widehat{MSE}_i^{(m)}$ is the bootstrap MSE estimate obtained in MC sample m. We define the average empirical coverage of the prediction

interval by $CR = M^{-1}\sum_{m=1}^{M} D^{-1}\sum_{i=1}^{D} I[\bar{y}_i^{(m)} \in CI_i^{(m)}]$, where $CI_i^{(m)}$ is the normal theory prediction interval defined in Section 3.1 and obtained in MC simulation m. The relative bias of the bootstrap MSE estimator is -0.27%. The coverage rate is 93.00%. The bootstrap appears to provide a reasonable measure of the uncertainty of the predictor.

5.4. Evaluating the Bayesian Posterior MSE

Now, we want to test the quality of the posterior MSE as a measure of the MSE of the Bayesian predictor. We consider the simulation configuration with $\rho = 0$. For each MC simulation, we use the procedure described in Section 3.2 to obtain the posterior MSE. We define the average relative bias of the posterior MSE as an estimator of the MSE of the Bayesian predictor. For Stan, the relative bias is defined by

$$RB_{Stan} = \frac{MSE_{Stan} - MSE_5}{MSE_5},$$

where $\text{MSE}_5 = \frac{1}{D}\sum_{i=1}^{D}\frac{1}{M}\sum_{m=1}^{M}(\hat{\bar{y}}_{N_i}^{(5,m)} - \bar{y}_i^{(m)})^2$,
$MSE_{Stan} = M^{-1}\sum_{m=1}^{M} D^{-1}\sum_{i=1}^{D}\widehat{MSE}_i^{B(m),5}$, and $\widehat{MSE}_i^{B(m),5}$ is the posterior MSE (10) obtained in MC simulation m from Stan. Extreme values are removed before calculating the relative bias for Stan. The relative bias for INLA is defined similarly as

$$RB_{INLA} = \frac{MSE_{INLA} - MSE_6}{MSE_6},$$

where $\text{MSE}_6 = \frac{1}{D}\sum_{i=1}^{D}\frac{1}{M}\sum_{m=1}^{M}(\hat{\bar{y}}_{N_i}^{(6,m)} - \bar{y}_i^{(m)})^2$,
$MSE_{INLA} = M^{-1}\sum_{m=1}^{M} D^{-1}\sum_{i=1}^{D}\widehat{MSE}_i^{B(m),6}$, and $\widehat{MSE}_i^{B(m),6}$ is the posterior MSE (10) obtained in MC simulation m from INLA. We define the average empirical coverage of the prediction interval by $CR = M^{-1}\sum_{m=1}^{M} D^{-1}\sum_{i=1}^{D} I[\bar{y}_i^{(m)} \in CI_i^{B(m)}]$, where $CI_i^{B(m)}$ is the prediction interval defined in Section 3.2 and obtained in MC simulation m using either Stan or INLA. The relative biases of the posterior MSE estimators are 20.80% from Stan and 16.68% from INLA. When the extreme values are included, the relative bias for Stan is closer to 70.00%. The slight improvement in relative bias from INLA may occur because of the extreme values generated in a small number of MC samples from Stan. The coverage rate for Stan is 96.77%, and for INLA is 96.27%. Although the MC mean of the posterior mean square error is higher than the mean square error of the Bayesian predictor, the coverage rate is reasonable. These results remained stable across correlation values of 0.2 and 0.6 as well.

5.5. Conclusions for simulations

Based on the simulation, the frequentist procedure's bootstrap MSE estimator is a good approximation for the MSE of the frequentist predictor. The posterior MSE based on either Stan or INLA over-estimates the MSE of the Bayesian predictor in the simulations. The over-estimation from INLA is less severe than the overestimation based on Stan. This suggests that the frequentist and INLA procedures may give more reliable measures of uncertainty in this context.

The Bayesian procedure is easier to extend to more complex models, such as models that incorporate data for multiple years. Further, considering the coverage rate of the prediction intervals from the frequentist and Bayesian procedures, there is no important difference between the two. Therefore, we will apply both the frequentist and Bayesian methods for the unit-level one-inflated beta model to the Seat-Belt Survey data.

6. Application to Seat Belt Survey Data

We next apply the predictors defined in Section 3 to the seat-belt survey data. Let $i = 1, \ldots, 15$ index the 15 counties. Let $j = 1, \ldots, n_i$ index the road segments in the sample for county i. The response variable, y_{ij}, represents the proportion of drivers who are observed to be wearing a seat-belt on a road segment j of county i. The small area parameters of interest are the proportions of belted drivers in the 15 sampled counties. These small area parameters are defined formally in (4).

The two covariates are the two design variables that are used to select the road segments within counties. Use of the design variables as covariates essentially ensures that the sample design is non-informative for the specified model. The first covariate is the vehicle miles traveled (VMT) for the road segment. The VMT is a measure of the average annual vehicle miles traveled across the road segment. Because the raw VMT values are skewed right, we use the log of the VMT as the model covariate. Let $x_{1,ij}$ denoted the log VMT for road segment j in county i. The second covariate is the road type. The road type is a categorical covariate with three levels: local, primary, and secondary. Let R_{ij} denote the road type for a road segment j of county i, where $R_{ij} \in \{\text{Primary}, \text{Secondary}, \text{Local}\}$. We let $\boldsymbol{x}_{2,ij} = (I[R_{ij} = \text{Primary}], I[R_{ij} = \text{Secondary}])'$, where $I[]$ is the indicator function. With this definition of $\boldsymbol{x}_{2,ij}$, local roads form the baseline category. The covariates are available for every road segment in the population from the sampling frame.

6.1. Variable Selection

Since we build the analysis based on the unit-level one-inflated model, there are two parts, the binary and beta parts, that need to be estimated. For each part, the two possible model covariates are $x_{1,ij}$ (log VMT) and $\boldsymbol{x}_{2,ij}$ (Road Type). Currently, we do not know which covariates will be significant for the data model. Therefore, we will use step-wise selection to choose the best model. The step-wise selection process begins with a model that contains all variables (called the Full Model). Then, variables are removed if omitting the variable leads to a decrease in the AIC. For the seat-belt survey data, the full models for the beta and binary parts of the one-inflated beta model are given respectively by

$$\text{logit}(\mu_{ij}) = \beta_0 + x_{1,ij}\beta_1 + \boldsymbol{x}'_{2,ij}\boldsymbol{\beta}_2 + b_i$$

and

$$\text{logit}(p_{ij}) = \alpha_0 + x_{1,ij}\alpha_1 + \boldsymbol{x}'_{2,ij}\boldsymbol{\alpha}_2 + u_i.$$

We may not need to include both $x_{1,ij}$ and $\boldsymbol{x}_{2,ij}$ in each model part. We apply the step-wise AIC method to the beta model part and binary part separately. We also use the cAIC for the binary part. We do not use the cAIC for the beta part because the `glmmmTMB` R package does not support the cAIC. The purpose of using step-wise variable selection is to identify a more parsimonious model that contains the important covariates.

Table 3. Output from Step-AIC for Beta model selection. A $*$ indicates the smallest AIC.

Variables	Model AIC
`Vehicle.Miles.Traveled` and `Road.Type`	-291.51 *
`Vehicle.Miles.Traveled`	-290.34
`Road.Type`	-280.73
`NONE`	-270.78

Table 4. Output from Step-cAIC for Binary model selection. A $*$ indicates the smallest cAIC or AIC.

Variables	Model cAIC	Model AIC
`Vehicle.Miles.Traveled` and `Road.Type`	44.85*	57.3
`Vehicle.Miles.Traveled`	45.92	53.4 *
`Road.Type`	50.98	59.0

Table 3 presents the results of the step-wise AIC process for `glmmTMB`, and table 4 presents the results of the step-wise cAIC and AIC processes for `glmer`. Based on the step AIC method, we find the `Vehicle.Miles.Traveled` and the `Road.Type` should be kept in beta part. We found that the cAIC was not a useful measure for the binary part. The problem is that the degrees of freedom for cAIC are always zero. Therefore, cAIC always decreases when more parameters are added to the `glmer` model. Therefore, we will select the model based on the AIC results. Using the AIC criterion, only the `Vehicle.Miles.Traveled` should be maintained in the binary part. This result will be used for both Bayesian and frequentist analyses.

6.2. Estimation of Fixed Model Parameters Using Bayesian and Frequentist Approaches

Table 5 contains the estimates of the fixed model parameters and corresponding measures of uncertainty. The Bayesian estimate is the posterior mean, and the Bayesian standard deviation is the posterior standard deviation for both Stan and INLA. The frequentist standard errors are based on large sample theory for all parameters except for the variance components, where we use the bootstrap. In both frequentist and Bayesian (Stan and INLA) frameworks, based on β_1, we infer that the seat-belt use rate increases with log(VMT), and log(VMT) is a significant predictor of the seat-belt use rate. Further, the seat-belt use rates are the highest on primary roads, followed by secondary roads, which can be observed from β_2 and β_3. The log(VMT) is inversely related to the probability of observing a 1, according to both Bayesian and frequentist procedures. The log(VMT) is significant in the model for the binary component. The negative association between log(VMT) and the probability of observing a 1 seems counterintuitive at first. This negative association occurs because seat-belt use rates of 1 are more likely to occur when the traffic volume is lower. Inferences based on the Bayesian and frequentist inference procedures are similar for most parameters,

except for the random effect variances, where the estimates based on the Bayesian method exceed those of the frequentist procedure.

Most of the parameter estimates are stable, in the sense that the estimates exceed twice the standard error in magnitude. Exceptions to this are the random effect variances, which do not differ significantly from zero. This survey data has only 15 counties, and the model has 9 parameters. Estimation of random effect variances with few degrees of freedom is difficult.

The scale reduction factors (Rhat) from Stan are presented in the last column of Table 5. All scale reduction factors are below 1.01, indicating that the chains have converged. Scale reduction factors for random effects are included in the histogram in Section 5 of the supplement. As INLA does not use MCMC, examining convergence for INLA is not necessary.

Table 5. Estimates and Standard Errors for Fixed Parameters

	Estimate			Standard Error			Rhat
Parameters	Frequentist	Stan-Bayes	INLA-Bayes	Frequentist	Stan-Bayes	INLA-Bayes	Stan-Bayes
β_0	1.10	1.13	1.13	0.40	0.41	0.41	1.00
β_1	0.19	0.19	0.19	0.06	0.06	0.06	1.00
β_2	0.47	0.47	0.48	0.24	0.25	0.24	1.00
β_3	0.12	0.11	0.11	0.18	0.19	0.18	1.00
ϕ	42.50	42.70	41.68	9.07	7.96	7.71	1.00
α_0	2.00	2.60	2.71	1.95	2.82	2.11	1.00
α_1	-0.62	-0.86	-0.82	0.31	0.41	0.33	1.00
σ_b	0.08	0.33	0.12	0.06	0.12	0.09	1.00
σ_u	0.12	1.81	1.95	1.66	2.50	2.13	1.00

To validate the goodness of fit of the model, we calculated posterior predictive p-values for Stan and INLA. For most statistics, the posterior predictive p-values are in the range of 0.05 to 0.95, indicating that the data are compatible with the model. For the skewness and kurtosis, the p-values are outside this range, suggesting that improvements to the model could be explored for the purpose of describing higher order moments of the data. Nonetheless, on the whole, the the posterior predictive p-values indicate that the dominant features of the model are compatible with the data.

6.3. Prediction Results

Next, we use the Bayesian (Stan and INLA) and frequentist procedures to obtain predictors of seat-belt use rates among drivers and corresponding measures of uncertainty for the 15 counties. We will compare the model-based predictors to the direct estimators, where the direct estimators are calculated with the sampling weights. Specifically, the direct estimators and the corresponding standard errors are calculated as if the design were probability proportional to size with replacement (PPSWR) within counties. In summary, we have four predictors for each county: (1) the direct estimator, (2) the model-based frequentist predictor, (3) the model-based Stan-Bayesian predictor, and (4) the model-based INLA-Bayesian predictor. We also have three measures of uncertainty for each county: (1) the direct standard error, (2) the bootstrap root-mean-square error, (3) the Stan-Bayesian posterior standard deviation, and (4) the INLA-Bayesian posterior standard deviation. We compare these alternative predictors and measures of uncertainty in this section.

Table 6 and Figures 6-8 allow a comparison of the direct estimator to the Bayesian and frequentist predictors. Table 6 contains the direct estimators and model-based predictors for the 15 counties, along with the corresponding measures of uncertainty.

The left panel of Figure 7 contains a plot of the frequentist predictors on the vertical axis against the direct estimators on the horizontal axis. The right panel of Figure 7 plots the bootstrap root-mean-square error estimates against the standard errors of the direct estimators. The solid lines are 45-degree lines through the origin. Figure 8 is analogous to Figure 7, with the Stan-Bayesian predictors and posterior root-mean-square errors in place of the frequentist estimates and standard errors. Also, figure 9 is analogous to Figure 7, with the INLA-Bayesian predictors and posterior root-mean-square errors in place of the frequentist estimates and standard errors.

Table 6. Comparison between direct estimators, frequentist predictors based on the unit-level one-inflated beta model, and Bayesian predictors based on the unit-level one-inflated beta model

	Estimator				Root MSE			
	Direct	Model-Based			Direct	Model-Based		
County ID	Estimator	Frequentist	Stan-Bayesian	INLA-Bayesian	Direct Model	Frequentist	Stan-Bayesian	INLA-Bayesian
1	0.9216	0.9287	0.9177	0.9174	0.0102	0.0276	0.0209	0.0213
2	0.9415	0.9296	0.9299	0.9292	0.0106	0.0217	0.0158	0.0157
3	0.8946	0.9108	0.9115	0.9095	0.0498	0.0214	0.0205	0.0208
4	0.9604	0.9416	0.9395	0.9390	0.0136	0.0295	0.0184	0.0184
5	0.9707	0.9350	0.9442	0.9415	0.0164	0.0245	0.0171	0.0176
6	0.8867	0.9080	0.9094	0.9076	0.0515	0.0271	0.0248	0.0245
7	0.9031	0.9142	0.9045	0.9046	0.0136	0.0244	0.0226	0.0231
8	0.9724	0.9484	0.9543	0.9541	0.0037	0.0207	0.0128	0.0130
9	0.9737	0.9430	0.9476	0.9471	0.0078	0.0191	0.0149	0.0149
10	0.9543	0.9417	0.9442	0.9428	0.0249	0.0221	0.0163	0.0163
11	0.9631	0.9493	0.9546	0.9542	0.0151	0.0193	0.0108	0.0109
12	0.9098	0.9058	0.908	0.9069	0.0165	0.0228	0.0246	0.0244
13	0.9481	0.9312	0.9345	0.9340	0.0073	0.0203	0.0149	0.0151
14	0.9225	0.9278	0.9263	0.9256	0.0195	0.0207	0.0164	0.0166
15	0.9737	0.9367	0.9465	0.9439	0.0123	0.0245	0.0156	0.0170

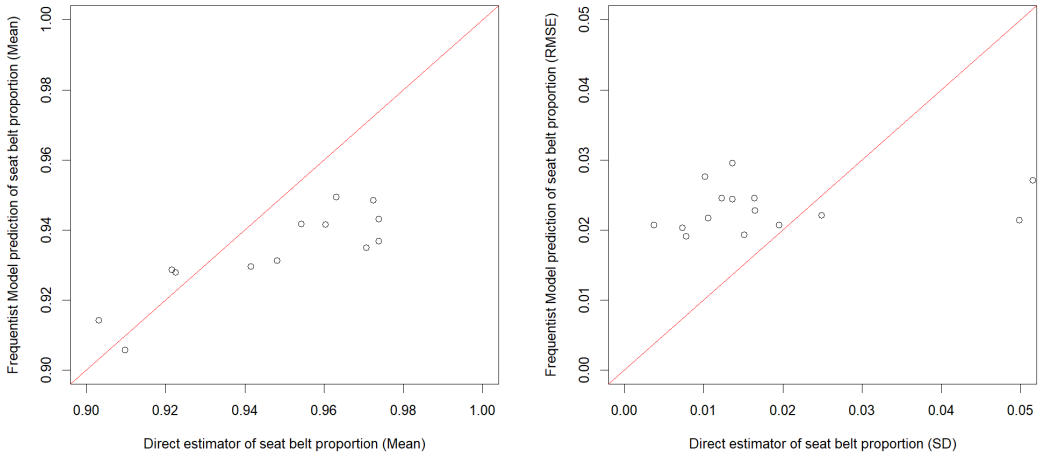

Figure 6. Comparison between direct estimation and frequentist model estimation for area-level seat-belt use rates. The red line is the (0,1) line.

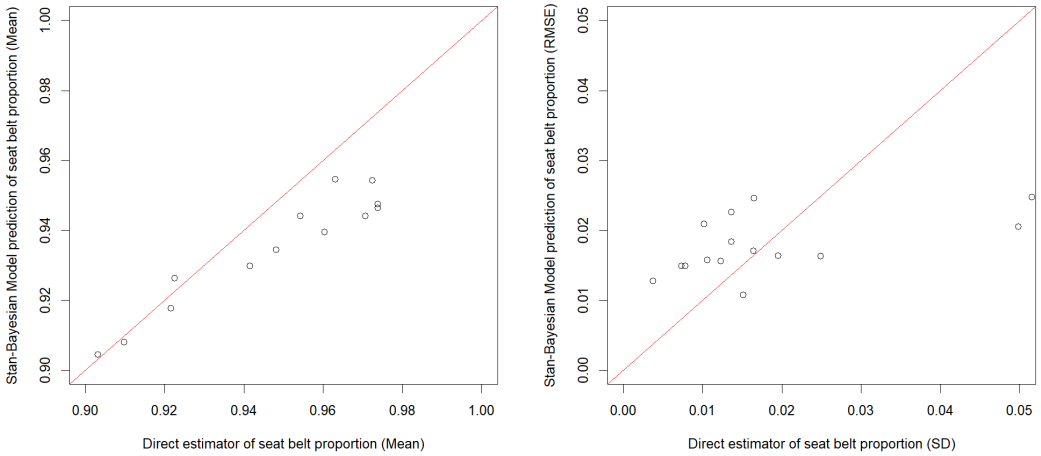

Figure 7. Comparison between direct estimation and Bayesian model estimation for area-level seat-belt use rates. The red line is the (0,1) line.

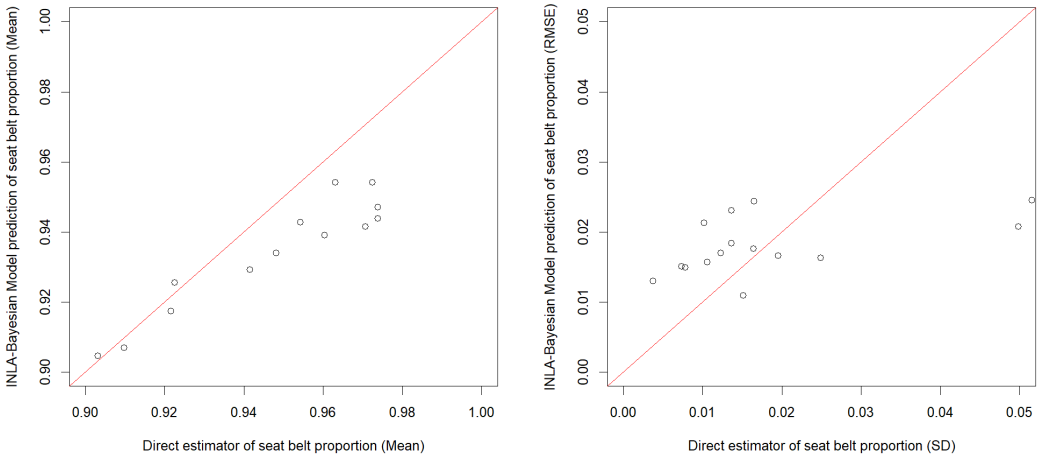

Figure 8. Comparison between direct estimation and INLA-Bayesian model estimation for area-level seat-belt use rates. The red line is the (0,1) line.

Table 6 and Figures 6-8 reveal several interesting trends. Based on both Bayesian and frequentist procedures, prediction has the anticipated effect of regression toward the mean. Direct estimators that are below average are increased, and direct estimators that are above average are decreased. This pattern is observed in the left panels of Figures 6 to 8. In our interpretation, neither the Bayesian nor frequentist predictors exhibit evidence of systematic bias. The predominant pattern in Figures 6 to 8 is regression toward the mean.

The properties of the measures of uncertainty are similar to the properties of the predictors in the sense that modeling shrinks the measures of uncertainty toward the overall average. The model-based root mean square errors vary less around the mean than the direct standard errors. A positive effect of this is that modeling has

the effect of reducing direct standard errors that are extremely large. The model leads to a substantial reduction in RMSE for the two counties, as highlighted in Table 6, where the direct standard deviations are exceptionally large. However, the model also increases direct standard errors that are unusually small. The property that the model-based root means square errors can exceed the direct standard errors is not necessarily bad. To understand this phenomenon, it is important to make a distinction between the true mean square error and the estimated mean square error. In the data analysis, we only have access to the estimated variances of the direct estimators and the estimated mean square errors of the model-based predictors. The true variances of the direct estimators and the true mean square errors of the model-based predictors are unknown. The direct standard errors are based on small sample sizes and can therefore be unstable. In our simulation studies, we found that the direct standard errors estimating the standard deviations of the direct estimators are often too small and can be below the estimated model-based root MSE's. This property is illustrated through the boxplots in Section 2 of the supplement, which display the ratios of the estimated RMSE's based on the model to the estimated standard deviations of the direct estimators in the simulation study. We think that modeling smooths the direct standard errors and thereby leads to a more reliable measure of uncertainty.

In the context of interpreting the measures of uncertainty in the data analysis, it is important to clarify that the sampling units in the survey are road segments. The relevant sample size for calculating variances is the number of road segments selected in a county. In this application, the number of road segments per county is 5 for all but one county where it is 14. Because the sample sizes are nearly constant across the counties, the mean square errors are stable across counties, and a plot of estimated variance against sample size is uninteresting for this application.

7. Discussion

We built a unit-level one-inflated beta model to obtain small area estimates of seat-belt use rates in 2018 among drivers. Model covariates include the road type and the vehicle miles traveled. The small areas are the 15 counties where the sample data were collected.

We motivated our interest in this model through an exploratory data analysis. The observed proportions have a skewed distribution on the interval $(0, 1]$. A linear association between the logits of the observed proportions and the log VMT is observed.

We compared the unit-level one-inflated beta model with the EBLUP for the linear model and the Gaussian one-inflated model through simulation. These two models do not preserve the parameter space for the seat-belt use rates. This problem is reflected in relatively high mean square errors in the simulation. We therefore do not apply these two alternative methods to the data.

In the real data analysis, we used the step AIC method to decide which covariates to include in the beta and binary parts of the unit-level one-inflated model. The vehicle miles traveled is important for both parts, and road type is important in the beta part.

In this application, we have decided only to produce estimates for the 15 sampled counties. Constructing estimates for the remaining 84 counties would require constructing predictions for counties with zero sampling units. This would require constructing predictions for 84 counties using only the regression component of the model because we cannot construct a predicted random effect for counties with zero

sample size. Use of only the regression component for a high percentage of counties seems ill-advised because the predictions would hinge on model assumptions that cannot be validated for counties with sample size of zero. Because we only construct estimates for the 15 sampled counties, the usual benchmarking operation, which forces the sum of the small area predictors to equal the direct estimate for the overall population, is inappropriate. Nonetheless, the estimates for the 15 sampled counties are of inherent interest in the context of the application.

The predictors based on the frequentist and Bayesian procedures are similar. More importantly, the Bayesian procedure is easier to extend to more complex model structures, such as models that incorporate data for multiple time-points. In the future, we will consider estimating the seat-belt use rates for multiple years. We will extend this unit-level one-inflated beta model to see whether there is a temporal trend in the behavior of the estimated regression coefficients across years. In this direction, we will use the Bayesian method, particularly INLA. INLA produces more stable results than Stan in the simulation study and is straightforward to extend to more complex models.

References

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

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

[3] M. E. Brooks, K. Kristensen, K. J. Van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Machler, and B. M. Bolker, *glmmtmb balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling*, The R journal, 9 (2017), pp. 378–400.

[4] H. Chandra and U. Sud, *Small area estimation for zero-inflated data*, Communications in Statistics-Simulation and Computation, 41 (2012), pp. 632–643.

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

[6] E. Fabrizi, *Inflated beta models for poverty mapping. an application integrating survey and remote sensing data*, 5 2022. Presentation at the SAE 2022 Conference, College Park, MD.

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

[8] D. Fletcher, D. MacKenzie, and E. Villouta, *Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression*, Environmental and ecological statistics, 12 (2005), pp. 45–54.

[9] A. Gelman, *Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)*, (2006).

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

[11] R. Janicki, *Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates*, Communications in Statistics-Theory and Methods, 49 (2020), pp. 2264–2284.

[12] J. Jiang and P. Lahiri, *Mixed model prediction and small area estimation*, Test, 15 (2006), pp. 1–96.

[13] M. Jumiartanti, I. Indahwati, and A. Kurnia, *Zero inflated beta model in small area estimation to estimate poverty rates on village level in langsa municipality*, Repositories-Dept. of Statistics, IPB University, (2017), pp. 812–819.
[14] K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. Bell, *Tmb: automatic differentiation and laplace approximation*, arXiv preprint arXiv:1509.00660, (2015).
[15] D. Lambert, *Zero-inflated poisson regression, with an application to defects in manufacturing*, Technometrics, 34 (1992), pp. 1–14.
[16] X. Lyu, E. J. Berg, and H. Hofmann, *Empirical Bayes small area prediction under a zero-inflated lognormal model with correlated random area effects*, Biometrical Journal, 62 (2020), pp. 1859–1878.
[17] P. McCullagh and J. Nelder, *Binary data*, in Generalized linear models, Springer, 1989, pp. 98–148.
[18] I. Molina, B. Nandram, and J. Rao, *Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach*, The Annals of Applied Statistics, 8 (2014), pp. 852–885.
[19] D. Pfeffermann, *New important developments in small area estimation*, Statistical Science, 28 (2013), pp. 40–68.
[20] D. Pfeffermann, B. Terryn, and F. A. Moura, *Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries*, Survey Methodology, 34 (2008), pp. 235–249.
[21] J. N. Rao and I. Molina, *Small area estimation*, John Wiley & Sons, 2015.
[22] D. F. Souza and F. A. Moura, *Multivariate beta regression with application in small area estimation*, Journal of Official Statistics, 32 (2016), pp. 747–768.
[23] A. H. Welsh, R. B. Cunningham, C. F. Donnelly, and D. B. Lindenmayer, *Modelling the abundance of rare species: statistical models for counts with extra zeros*, Ecological Modelling, 88 (1996), pp. 297–308.
[24] J. Wieczorek, C. Nugent, and S. Hawala, *A Bayesian zero-one inflated beta model for small area shrinkage estimation*, in Proceedings of the 2012 Joint Statistical Meetings, American Statistical Association, Alexandria, VA, 2012.