HIERARCHICAL MODELS FOR SMALL AREA ESTIMATION USING ZERO-INFLATED FOREST INVENTORY VARIABLES: COMPARISON AND IMPLEMENTATION

By

Grayson W. White

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics—Master of Science

2025

ABSTRACT

National Forest Inventory (NFI) data are typically limited to sparse networks of sample locations due to cost constraints. While traditional design-based estimators provide reliable forest parameter estimates for large areas, there is increasing interest in model-based small area estimation (SAE) methods to improve precision for smaller spatial, temporal, or biophysical domains. SAE methods can be broadly categorized into area- and unit-level models, with unit-level models offering greater flexibility—making them the focus of this study. Ensuring valid inference requires satisfying model distributional assumptions, which is particularly challenging for NFI variables that exhibit positive support and zero inflation, such as forest biomass, carbon, and volume. Here, we evaluate a class of two-stage unit-level hierarchical Bayesian models for estimating forest biomass at the county-level in Washington and Nevada, United States. We compare these models to simpler Bayesian single-stage and two-stage frequentist approaches. To assess estimator performance, we employ simulated populations and cross-validation techniques. Results indicate that small area estimators that incorporate a two-stage approach to account for zero inflation, county-specific random intercepts and residual variances, and spatial random effects provide the most reliable county-level estimates. Additionally, findings suggest that unit-level cross-validation within the training dataset is as effective as area-level validation using simulated populations for model selection. We also illustrate the usefulness of simulated populations for better assessing qualities of the various estimators considered.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

National Forest Inventories (NFIs) play a critical role in collecting data and monitoring forest trends to assess resource availability, health, composition, and other economic and ecological attributes across various spatial scales within a given country. In the United States, the NFI is conducted by the United States Department of Agriculture (USDA) Forest Service through the Forest Inventory and Analysis (FIA) Program. Traditionally, NFIs such as FIA have been designed to provide precise estimates at broader spatial scales, such as state-level assessments of forest attributes like timber volume and biomass. However, there is growing national interest, along with increased funding, in obtaining more precise biomass estimates at finer spatial scales, such as the county-level (Prisley et al., 2021; Wiener et al., 2021; U.S. Senate, 2023). This rising demand, coupled with the widespread availability of high-resolution remote sensing data, has prompted researchers to develop and apply innovative small area estimation (SAE) methods that integrate FIA data with remote sensing products (Cao et al., 2022; May et al., 2023; Finley et al., 2024).

Despite the wide variety of SAE methods, they can generally be categorized into two main approaches: area-level and unit-level methods (Rao and Molina, 2015). Both aim to estimate the same parameter of interest but differ significantly in their use of data. In area-level modeling, survey unit response variable measurements are aggregated at each area. These aggregates are referred to as direct estimates and are typically generated using a design-based estimator. Direct estimates are then set as the area-level response variable in a regression model that might include area-level summaries of predictor variables and structured random effects. The goal of areal models is to use sources of auxiliary information to smooth noisy direct estimates. In contrast, unit-level approaches retain response variable measurements at the individual unit level. Set as the response variable, these unit-level measurements are coupled with spatial and/or temporally aligned predictor variables, possibly along with structured random effects, in a predictive model. This predictive model is then used to predict for all unobserved units. Finally, predictions are aggregated to any user-defined area of interest.

The advantages and trade-offs of SAE approaches have only begun to be explored in the forest inventory literature. Area-level modeling often benefits from a more linear relationship between response and predictor variables and does not require precise plot locations, which is particularly useful given the often confidentiality of NFI plot data. However, aggregation leads to data loss, limits the ability to model fine-scale (i.e., unit-level) relationships, precludes delineation of new areas of interest after model fitting, and imposes statistical assumptions that might be difficult to justify. For example, in the classical Fay-Herriot model, within-area variance of the direct estimate is assumed to be fixed and known, although in practice it is estimated from limited data. These variances enter the area-level model through a random effect, often without strong theoretical justification or consistency between the two inferential paradigms. In contrast, unit-level approaches leverage precise plot locations to model fine-scale spatial relationships more effectively, making them particularly valuable when such data are available.

Unlike design-based estimators, which are determined by the sampling design, inference from model-based estimators relies entirely on the selection of an appropriate model. Consequently, we must take special care when specifying SAE models and conduct rigorous model checking. One of the most effective ways to assess SAE models is through simulated populations that closely resemble the true, but only partially observed, population of interest. Simulated populations allow us to explore how inference varies under different conditions (e.g., varying sample sizes) and to compare our estimates against "true" values. To ensure a meaningful evaluation, we must generate these simulated populations using methods that are not similar to the models we hope to assess.

Beyond assessment using simulated populations, we can evaluate models through cross-validation using observed data (e.g., leave-one-out or k-fold cross-validation). However, in SAE studies, the primary parameters of interest exist at the area level. Unit-level models can be assessed using cross-validation at the unit level (i.e., iteratively holdout one or more observations, predict for those holdout, and compare the predictions to the holdout true values); however, we must be careful to verify the unit-level assessments align with how well the estimator performs once predictions are aggregated to the desired areas of interest.

This study evaluates a range of unit-level SAE approaches for estimating average biomass at the county level in Washington and Nevada. A key challenge in this context arises when estimating biomass across areas with a mix of forest and non-forest landcover. Specifically, biomass values exhibit a mixture of continuous positive values and true zeroes, a phenomenon referred to here as "zero-inflation." While the term zero-inflation is commonly used in the statistical literature to describe a discrete distribution with an excessive number of zeros, in this case, biomass follows a continuous distribution with an additional zero component. Various model-based approaches have been developed to address zero-inflation in SAE. Notably, Pfeffermann et al. (2008) introduced a two-stage mixture model to account for zero-inflation in the response variable, exploring both frequentist and Bayesian modes of inference. Their findings suggest that mean squared error (MSE) estimation is more straightforward in the Bayesian paradigm due to the advantages of Markov chain Monte Carlo (MCMC) simulation in uncertainty propagation across the model stages. Expanding on this work, Chandra and Sud (2012) applied the same two-stage model in a frequentist setting and introduced a parametric bootstrap-based MSE estimator.

In forest inventory applications, zero-inflation has received relatively limited attention. Finley et al. (2011) developed a two-stage model for zero-inflation in continuous forest attributes such as biomass, volume, and age, employing a hierarchical Bayesian framework with Gaussian process-based spatial random effects. Their approach enables unit-level predictions of forest attributes along with uncertainty quantification, though they did not directly produce small area estimates. More recently, White et al. (2025) applied the zero-inflated SAE model from Chandra and Sud (2012) to FIA data in Nevada, generating county-level biomass estimates. Their study compared the zero-inflated estimator to other commonly used small area estimators, including estimators based on the Battese-Harter-Fuller unit-level model and the Fay-Herriot area-level model (Fay and Herriot, 1979; Battese et al., 1988). Their simulation results indicate the zero-inflated estimator improves point estimates and produces competitive MSE estimates, though further refinements remain possible (see Figure 2 in White et al., 2025).

In this study, we compare and extend model-based SAE approaches that account for zero-

inflation, applying them to FIA data and remote sensing products for Washington and Nevada, as described in Section 2.1. Specifically, we evaluate nine model-based approaches, including the zero-inflated estimator from Chandra and Sud (2012) and eight hierarchical Bayesian estimators of increasing complexity. These Bayesian estimators include both single-stage models that do not explicitly address zero-inflation and two-stage models designed to account for a preponderance of zeros. With state counties defining our small areas of interest, we investigate the effects of incorporating county-varying intercepts, county-varying coefficients, county-specific residual variances, and spatially-varying intercepts. Section 2.2 provides further details on these models. Rather than defaulting to the most complex estimator, we sequentially introduce additional model components and evaluate their impact on estimate qualities. This approach allows us to identify when added complexity improves estimation and when simpler models suffice. Relative to existing literature, Finley et al. (2011) considered spatial effects but did not include county-specific terms, while Pfeffermann et al. (2008) and Chandra and Sud (2012) did not incorporate spatial dependencies.

To evaluate these nine estimators, we conduct a simulation study based on the methodology of White et al. (2024a), detailed in Section 2.6. Chapter 3 presents the results of the simulation study and applies the estimators to FIA data, with cross-validation used to assess model performance at the unit level. Finally, we summarize our findings, discuss their implications, and suggest directions for future research in Chapter 4.

CHAPTER 2

METHODS

2.1 Data

The motivating data are from the USDA Forest Service FIA program and comprise inventory plot measurements of live aboveground tree biomass density (Mg/ha). These data were drawn from the most current sampled measurement for each plot in the FIA database downloaded on February 8, 2023 for the states of Washington and Nevada (Burrill et al., 2023). Washington was selected because it has large differences in biomass across counties, ranging from massive biomass densities on the Olympic Peninsula to near zero biomass in counties east of the Cascade Mountain Range. Nevada was selected because it has a fairly unique distribution of forest biomass, where much of the state's arid environment has little to no biomass which is punctuated with sky islands where there is non-zero forest biomass. Each state has approximately ten years of FIA data, ending in year 2019, that were derived from a panel of plots measured annually across a systematic sample of hexagons approximately 2,500 hectares in size. Biomass values were from live trees only and include all trees 1.0 inch diameter and greater.

Estimators and simulations, described in Sections 2.2 and 2.6, respectively, were informed using five auxiliary variables: National Land Cover Dataset Analytical Tree Canopy Cover 2016 (hereafter `tcc`); LANDFIRE 2010 Digital Elevation Model (hereafter `elev`); US Geological Survey Terrain Ruggedness Index (hereafter `tri`); PRISM mean annual precipitation, 30yr normals (1991-2020) (hereafter `ppt`); and LANDFIRE 2014 tree/non-tree lifeform mask (hereafter `tnt`) (Daly et al., 2002; Rollins, 2009; Yang et al., 2018; Picotte et al., 2019; U.S. Geological Survey, 2019). The `tcc` variable is a measure of average tree canopy cover in a given pixel, the `elev` variable gives the elevation at a given pixel, the `tri` variable gives the terrain ruggedness at a given pixel, the `ppt` variable is a measure of average precipitation at a given pixel over 30 years, and the `tnt` variable is a binary variable distinguishing between pixels with and without trees. These auxiliary variables were resampled to 90 meter resolution and available wall-to-wall in both states. At locations with FIA plots these variables are matched with the corresponding plot and then used as predictors in

the models' regression components and to inform simulated population generation.

2.1.1 Variable selection

Given the five auxiliary variables available to us, we had to decide when and where to use the auxiliary variables in our analyses. Our analyses spanned the generation of simulated populations for the states of Washington and Nevada, the fitting of model-based estimators on these simulated populations, and the fitting of model-based estimators on real data collected by FIA.

We first chose auxiliary variables for the simulated population generation. In both the states of Washington and Nevada, we believe the `tcc` and `elev` variables to be central to the underlying process which drive the quantity of biomass in a given location. Further, in Nevada, we believe the `tri` variable to have importance due to how biomass occurs on sky islands in Nevada, so we included this variable for the simulated population generation in Nevada. In Washington, we believe the `ppt` variable to be indicative of biomass quantities, as Washington gets substantially more precipitation West of the Cascade mountain range. For both states, we also chose to stratify the population by the `tnt` variable to ensure imputation occurs separately for treed and non-treed strata.

We next chose auxiliary variables for the model-based estimators. For simplicity, we used the same predictors on the simulated and FIA data. We restricted our analyses in each state to only include variables that we believe impact the data generation process, and followed suit with how we generated the simulated population. However, we initially did not include the `tnt` variable in our model-based approaches. We found that this approach was sufficient in Nevada, but in Washington the Bernoulli models produced imprecise predictions without the inclusion of the `tnt` variable. Therefore, in Washington we included the `tnt` variable for all Bernoulli models.

A description of where these variables were used in different model types and simulated population generation is shown in Table 2.1.

2.2 Model-based estimation

Nine candidate model-based estimators are used to estimate the average biomass at the county level across Nevada and Washington. The first estimator follows a frequentist mode of inference

Predictor	Gaussian	Bernoulli	Simulated population
tcc	WA, NV	WA, NV	WA, NV
elev	WA, NV	WA, NV	WA, NV
tri	NV	NV	NV
ppt	WA	WA	WA
tnt	none	WA	WA, NV

Table 2.1 Predictor variables used to inform estimators and for generating simulated populations.

and is constructed using a two-stage regression. The eight additional estimators use a Bayesian mode of inference and are constructed using one- and two-stage regressions. A brief description of the candidate estimators is provided in Table 2.2.

Estimator	Description
F ZI CVI	A frequentist two-stage estimator. The first stage model is a generalized linear mixed model with a county-varying intercept, and the second stage model is a linear mixed model with a county-varying intercept.
B CVI	A Bayesian single-stage estimator based on a linear mixed model with a county-varying intercept.
B CVC	The same as B CVI, but with county-varying coefficients.
B ZI CVI	A Bayesian two-stage estimator. The first stage model is a generalized linear mixed model with a county-varying intercept, and the second stage model is a linear mixed model with a county-varying intercept.
B ZI CVC	The same as B ZI CVI, but with county-varying coefficients.
B ZI CVI CRV	The same as B ZI CVI, but with county-specific residual variances.
B ZI CVC CRV	The same as B ZI CVC, but with county-specific residual variances.
B ZI CVI SVI CRV	The same as B ZI CVI CRV, but with an added spatial random effect, modeled as a Nearest Neighbor Gaussian process (NNGP) on the intercept.
B ZI CVC SVI CRV	The same as B ZI CVC CRV, but with an added spatial random effect, modeled as a NNGP on the intercept.

Table 2.2 Description of the candidate models considered for estimating county-level forest biomass. Abbreviations are: frequentist (F); Bayesian (B); zero-inflated (ZI); county-varying intercept (CVI); county-varying coefficients (CVC); county-specific residual variance (CRV); space-varying intercept (SVI).

Models are fit at the unit level, which sets plot-level biomass (Mg/ha) as the response variable. To better meet subsequent models' assumption of normally distributed residuals and to ensure

positive support for predictions, models are fit to a transformed response variable. Often in such settings the logarithm is a natural choice; however, we found the logarithm to be too strong, resulting in a skewed response distribution once zeros were accommodated. Due to the logarithm's undesirable strength in this application, we prefer root function transformations that were chosen to be state specific, with stronger roots for high biomass states and weaker roots for low to moderate biomass states. Specifically, we used a fourth and square root transformation for Washington and Nevada, respectively.

Despite our models being fit at the unit level, our inferential goal is to estimate average biomass per hectare at the county level, which we denote μ_j where j indexes county within a given state. In subsequent sections we present the candidate unit-level models and how each is used to estimate the small area parameters of interest.

2.3 Models for frequentist two-stage estimation

Here we introduce the models used to develop a frequentist two-stage approach to estimation (F ZI CVI). The resultant estimator and its MSE estimator were introduced by Chandra and Sud (2012) and later applied to forest inventory data by White et al. (2025) with promising results in the state of Nevada. At generic spatial location ℓ the transformed non-zero forest biomass is modeled as

$$y(\ell) = \beta_0 + \tilde{\beta}_0(\ell) + \boldsymbol{x}(\ell)^\top \boldsymbol{\beta} + \varepsilon(\ell), \tag{2.1}$$

where β_0 is the intercept, $\tilde{\beta}_0(\ell)$ is the county specific random effect with $\tilde{\beta}_0(\ell) = \tilde{\beta}_{0,j}$ when ℓ is in the jth county and with $\tilde{\beta}_{0,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\tilde{\beta}_0})$, $\boldsymbol{x}(\ell)$ is a $p \times 1$ vector of predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\varepsilon(\ell)$ is a residual error term with $\varepsilon(\ell) \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2)$.

Biomass presence and absence is modeled using a Bernoulli mixed model with a logit link function defined as

$$\log\left(\frac{p(\ell)}{1 - p(\ell)}\right) = \alpha_0 + \tilde{\alpha}_0(\ell) + \boldsymbol{v}(\ell)^\top \boldsymbol{\alpha}, \tag{2.2}$$

where $p(\ell)$ denotes the probability of non-zero response value at location ℓ, α_0 is the intercept, $\tilde{\alpha}_0(\ell)$ is the county specific random effect with $\tilde{\alpha}_0(\ell) = \tilde{\alpha}_{0,j}$ when ℓ is in the jth county and with

$\tilde{\alpha}_{0,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\tilde{\alpha}_0})$, $\boldsymbol{\nu}(\ell)$ is a $q \times 1$ vector of predictor variables, $\boldsymbol{\alpha}$ is a $q \times 1$ vector of regression coefficients.

2.4 Models for Bayesian estimation

Here we consider a class of hierarchical Bayesian models. These models are primarily two-stage models similar to the frequentist two-stage model developed in Section 2.3. For comparison, we also include some simpler, single-stage, models that do not explicitly accommodate excess zero values in the response.

The two-stage hierarchical model used in Finley et al. (2011) and adopted here is

$$\begin{aligned} y(\ell) \mid \text{parameters} &\sim \mathcal{N}\left(z(\ell)m(\ell),\ z(\ell)\tau_1^2 + (1 - z(\ell))\tau_2^2\right), \\ z(\ell) \mid \text{parameters} &\sim \mathcal{BER}\left(p(\ell)\right). \end{aligned} \tag{2.3}$$

The first level of hierarchy is the model for $z(\ell)$. In our case, we only consider two options for estimation of $z(\ell)$: first, setting $z(\ell)$ to 1; and second, using a linear mixed model with logit link function. In the second case, for simplicity, we only use one model for estimation of $z(\ell)$ across all Bayesian estimators. Here, a realization of $z(\ell)$ indicates whether or not the location ℓ is predicted to have biomass (1) or not (0). Then, we pass the realization of $z(\ell)$ into the second level of hierarchy where it determines the expression of the mean $m(\ell)$ and the associated variance term. In the subsequent development of models, $m(\ell)$ is a given linear mixed model with county-varying intercept (CVI), county-varying coefficients (CVC), or space-varying intercept (SVI) components (Table 2.2), resulting in four different model forms for $m(\ell)$ that we implement and compare. The residual variance is estimated through a parameter τ_1^2 when $z(\ell)$ is 1, and set to a small value via a constant, τ_2^2, when $z(\ell)$ is 0 (see Finley et al., 2011, Section 3 for details). For some candidate models, we allow for county-specific residual variance (CRV) terms via county-specific τ_1^2s (Table 2.2).

We now turn to the particular models that use the hierarchical structure given by Eq. (2.3). The simplest model we consider is the county-varying intercept (B CVI) model, which is the Bayesian equivalent to Eq. (2.1) and is defined as

$$y(\ell) = \beta_0 + \tilde{\beta}_0(\ell) + \boldsymbol{x}(\ell)^\top \boldsymbol{\beta} + \varepsilon(\ell), \tag{2.4}$$

with parameter and hyperparameter distributions defined as follows, $\beta_0 \sim \mathcal{N}(0, \sigma^2_{\beta_0})$, $\tilde{\beta}_0(\ell)$ is the county specific random effect, i.e., $\tilde{\beta}_0(\ell) = \tilde{\beta}_{0,j}$ and $\tilde{\beta}_0(\ell) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\tilde{\beta}_0})$ when ℓ is in the jth county, $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\beta}\boldsymbol{I})$ with $\boldsymbol{I}$ being the p-dimentional identify matrix, and $\varepsilon(\ell) \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2)$, $\sigma^2_{\tilde{\beta}_0} \sim \mathcal{IG}(a_{\sigma^2_{\tilde{\beta}_0}}, b_{\sigma^2_{\tilde{\beta}_0}})$, and $\tau^2 \sim \mathcal{IG}(a_{\tau^2}, b_{\tau^2})$. All hyperparameters were set to induce non-informative prior distributions.

Next, we consider the county-varying coefficient (B CVC) model that allows regression coefficients to vary by county. This model is defined as

$$y(\ell) = \beta_0 + \tilde{\beta}_0(\ell) + \boldsymbol{x}(\ell)^\top \left(\boldsymbol{\beta} + \tilde{\boldsymbol{\beta}}(\ell)\right) + \varepsilon(\ell), \tag{2.5}$$

where $\tilde{\boldsymbol{\beta}}(\ell) = (\tilde{\beta}_{1,j}, \tilde{\beta}_{2,j}, \ldots, \tilde{\beta}_{p,j})^\top$ with $\tilde{\beta}_{k,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\tilde{\beta}_{k,j}})$ and $\sigma^2_{\tilde{\beta}_{k,j}} \sim \mathcal{IG}(a_{\sigma^2_{\tilde{\beta}_{k,j}}}, b_{\sigma^2_{\tilde{\beta}_{k,j}}})$ for $k = 1, 2, \ldots, p$ when ℓ is in the jth county.

The B CVI and B CVC models defined above are used in our analyses both as a single-stage model (by setting $z(\ell) = 1$ in the hierarchical model) and in the two-stage setting. We are most interested in these simpler, single-stage, models for comparison with two-stage models as laid out in Table 2.2.

Turning now to the two-stage models. Each two-stage model introduced in this section uses the same first stage model; however, the framework we defined in Eq. (2.3) lends itself to a variety of different second stage models. We consider a range of second stage models.

The first stage model used for all two-stage models is the Bayesian equivalent to Eq. (2.2), a generalized linear mixed model with Bernoulli response and a county-varying intercept defined as

$$\log\left(\frac{p(\ell)}{1 - p(\ell)}\right) = \alpha_0 + \tilde{\alpha}_0(\ell) + \boldsymbol{v}(\ell)^\top \boldsymbol{\alpha}, \tag{2.6}$$

where $p(\ell)$ denotes the probability of non-zero response, and parameter and hyperparameter distributions are defined as follows, $\alpha_0 \sim \mathcal{N}(0, \sigma^2_{\alpha_0})$, $\tilde{\alpha}_0(\ell)$ is the county specific random effect, i.e., $\tilde{\alpha}_0(\ell) = \tilde{\alpha}_{0,j}$ and $\tilde{\alpha}_0(\ell) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\tilde{\alpha}_0})$ when ℓ is in the jth county, $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\alpha}\boldsymbol{I})$, and $\sigma^2_{\tilde{\alpha}_0} \sim \mathcal{IG}(a_{\sigma^2_{\tilde{\alpha}_0}}, b_{\sigma^2_{\tilde{\alpha}_0}})$. All hyperparameters were set to induce non-informative prior distributions.

Now, we can combine Eq. (2.6) with Eq. (2.4) and specify the first two-stage hierarchical model (B ZI CVI) as

$$y(\ell) = z(\ell)\left(\beta_0 + \tilde{\beta}_0(\ell) + \boldsymbol{x}(\ell)^\top \boldsymbol{\beta}\right) + z(\ell)\varepsilon_1(\ell) + (1 - z(\ell))\varepsilon_2(\ell), \tag{2.7}$$

where $\varepsilon_1(\ell) \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_1^2)$ and $\varepsilon_2(\ell) \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_2^2)$ with $\tau_1^2 \sim \mathcal{IG}(a_{\tau_1^2}, b_{\tau_1^2})$ and $\tau_2^2 = 0.00001$.

The hierarchical model specified in Eq. (2.7) is analogous to the frequentist two-stage model specified in Section 2.3. Notably, this hierarchical structure combines the Bernoulli model with the B CVI model to produce estimates that account for zero inflation.

We next extend Eq. (2.7) to include county-varying coefficients (B ZI CVC) defined as

$$y(\ell) = z(\ell)\left(\beta_0 + \tilde{\beta}_0(\ell) + \boldsymbol{x}(\ell)^\top \left(\boldsymbol{\beta} + \tilde{\boldsymbol{\beta}}(\ell)\right)\right) + z(\ell)\varepsilon_1(\ell) + (1 - z(\ell))\varepsilon_2(\ell). \tag{2.8}$$

In practice, it is common for residual variance to increase with increasing biomass, i.e., heteroscedasticity. Given disparity in forest density at the county-level, we might expect residual variance to vary across counties. For example, western counties in Washington have much more forest biomass than counties in eastern Washington, and if this disparity in biomass is not entirely captured by the predictors and random effects, then we would expect quite different residual variances. A county-specific residual variance term can help accommodate heteroscedasticity and improve county-level estimates; hence, we extend the B ZI CVI Eq. (2.7) and B ZI CVC Eq. (2.8) models with county-specific residual variance parameters. Specifically, the zero-inflated county-varying intercept model with county-specific residual variance (B ZI CVI CRV) and the corresponding county-varying coefficient model (B ZI CVC CRV) are defined analogous to Eq. (2.7) and Eq. (2.8) but with $\varepsilon_1(\ell) \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_{1,j}^2)$ with $\tau_{1,j}^2 \sim \mathcal{IG}(a_{\tau_1^2}, b_{\tau_1^2})$ when ℓ is in the jth county.

In addition to county scale differences in mean biomass and predictor variables' relationships with biomass, captured through county-varying intercepts and county-varying coefficients, respectively, we might expect to see smoothly-varying spatially structured changes in mean biomass caused by disturbance history, climate impacts, species composition, or any spatially dependent factors not captured by predictors variables. Such spatial changes in mean biomass can be accommodated via a space-varying intercept random effect. The models below extend the B ZI CVI CRV

and B ZI CVC CRV models to include such a space-varying intercept. Specifically, the B ZI CVI SVI CRV model is defined as

$$y(\ell) = z(\ell)\left(\beta_0 + \tilde{\beta}_0(\ell) + \boldsymbol{x}(\ell)^\top \boldsymbol{\beta} + w(\ell)\right) + z(\ell)\varepsilon_1(\ell) + (1 - z(\ell))\varepsilon_2(\ell), \tag{2.9}$$

where $w(\ell)$ is a spatial random effect that adjusts the intercept based on residual spatial dependence. Here we estimate $w(\ell)$ using a Gaussian Process approximation called the Nearest Neighbor Gaussian Process (NNGP; Datta et al. 2016; Finley et al. 2019) that provides substantial improvements in run time, with negligible differences in inference and prediction, compared to a model that uses a full Gaussian Process. In brief, for this specification, the vector of random effects collected over n locations $\boldsymbol{w} = (w(\ell_1), w(\ell_2), \ldots, w(\ell_n))^\top$ is distributed multivariate normal with mean zero and covariance matrix that captures the spatial dependence among random effects, i.e., $\boldsymbol{w} \sim \mathcal{MVN}(\boldsymbol{0}, \sigma_w^2 \boldsymbol{R}(\phi))$, where σ_w^2 is the spatial variance and $\boldsymbol{R}(\phi)$ is the NNGP-derived correlation matrix that depends on a spatial correlation function, which in our case is exponential, and decay parameter ϕ used to estimate the strength of correlation between any two locations. As with other variance parameters, we assume $\sigma_w^2 \sim \mathcal{IG}(a_{\sigma_w^2}, b_{\sigma_w^2})$ with hyperparameters set to induce a non-informative prior distribution. The spatial decay parameter is assumed to follow a uniform distribution with broad, non-informative, spatial support.

The last candidate model ZI CVC SVI CRV extends B ZI CVI SVI CRV Eq. (2.9) to include county-varying coefficients.

2.5 Model implementation and comparison

2.5.1 Frequentist two-stage estimator

Model parameters for the two-stage frequentist model (F ZI CVI) described in Section 2.3 were estimated using restricted maximum likelihood via the `R` (R Core Team, 2024) `saeczi` package (Yamamoto et al., 2025) which implements methods presented by Chandra and Sud (2012).

As described in Chapter 1, to generate estimates for a small area of interest, we predict biomass and probability of non-zero biomass using Eq. (2.1) and Eq. (2.2) respectively, over a fine grid of

prediction locations. For generic prediction location ℓ^* these predictions are

$$y(\ell^*) = \hat{\beta}_0 + \hat{\tilde{\beta}}_0(\ell^*) + \boldsymbol{x}(\ell^*)^\top \hat{\boldsymbol{\beta}}, \quad \text{and} \quad p(\ell^*) = \frac{\exp\left(\hat{\alpha}_0 + \hat{\tilde{\alpha}}_0(\ell^*) + \boldsymbol{v}(\ell^*)^\top \hat{\boldsymbol{\alpha}}\right)}{1 + \exp\left(\hat{\alpha}_0 + \hat{\tilde{\alpha}}_0(\ell^*) + \boldsymbol{v}(\ell^*)^\top \hat{\boldsymbol{\alpha}}\right)}, \tag{2.10}$$

where the $\hat{}$ indicates each parameter's maximum likelihood point estimate. The estimate for μ_j is then the average product of these predictions over the grid of prediction locations

$$\hat{\mu}_j = \frac{1}{n_j^*} \sum_{\ell \in U_j} g^{-1}\left(y(\ell^*)\right) p(\ell^*), \tag{2.11}$$

where $g^{-1}(\cdot)$ is the inverse of the transformation function used when fitting the model, allowing us to revert back to biomass scale, and U_j is the set of n^* prediction locations within the jth county. This estimator's MSE estimator comes from a parametric bootstrap introduced in Chandra and Sud (2012) and discussed and explored in White et al. (2025).

2.5.2 Bayesian estimators

Parameter inference for Bayesian models described in Section 2.4 was based on Markov chain Monte Carlo (MCMC) samples from posterior distributions. Gibbs and Metropolis Hastings algorithms were implemented in `C++` to efficiently sample from parameter posterior distributions. Code, additional information about the sampling algorithms, and example analyses using simulated data are given in Finley (2025) and a list of prior distributions and hyperparameter values is given in the Appendix. Posterior inference is based on M=3,000 post-convergence and thinned samples from three MCMC chains, i.e., 1,000 from each chain. We use convergence diagnostics and thinning rules outlined in Gelman et al. (2013).

Inference about biomass at prediction locations and subsequent county-level estimates for μ_j are based on posterior predictive distribution samples. All single- and two-stage models follow the same approach for predictive inference, which is based on composition sampling from each model's posterior predictive distribution. For example, using the B ZI CVI SVI CRV Eq. (2.9) model, for generic prediction location ℓ^* we generate M post-convergence and thinned samples one-for-one, first plugging in the sth MCMC sample of model parameters into the model's predictive distribution to generate a corresponding posterior predictive distribution sample. Specifically,

for $s = 1, 2, \ldots, M$ we draw a sample from the posterior predictive distribution for forest presence/absence

$$z^{(s)}(\ell^*) \sim \mathcal{RBER}\left(\frac{\exp\left(\alpha_0^{(s)} + \tilde{\alpha}_0^{(s)}(\ell^*) + \boldsymbol{\nu}(\ell^*)^\top \boldsymbol{\alpha}^{(s)}\right)}{1 + \exp\left(\alpha_0^{(s)} + \tilde{\alpha}_0^{(s)}(\ell^*) + \boldsymbol{\nu}(\ell^*)^\top \boldsymbol{\alpha}^{(s)}\right)}\right) \tag{2.12}$$

then, given $z^{(s)}(\ell^*)$, we draw from the posterior predictive distribution for transformed biomass

$$y^{(s)}(\ell^*) \sim \mathcal{RN}\Big(z^{(s)}(\ell^*)\left(\beta_0^{(s)} + \tilde{\beta}_0^{(s)}(\ell^*) + \boldsymbol{x}(\ell^*)^\top \boldsymbol{\beta}^{(s)} + w^{(s)}(\ell^*)\right), \\ z(\ell^*)^{(s)}\tau_1^{2(s)} + (1 - z(\ell^*)^{(s)})\tau_2^{2(s)}\Big), \tag{2.13}$$

where $\mathcal{RBER}$ and $\mathcal{RN}$ generate a random draw from a Bernoulli and normal distribution, respectively.

Given M posterior predictive samples from Eq. (2.13), we can draw corresponding samples for county-level means. Specifically, samples from the jth county's posterior predictive distribution are drawn one-for-one from

$$\mu_j^{(s)} = \frac{1}{n_j^*} \sum_{\ell \in U_j} g^{-1}\left(y^{(s)}(\ell^*)\right), \tag{2.14}$$

for $s = 1, 2, \ldots, M$.

For subsequent comparison and mapping, we calculate the mean and credible intervals using samples from the desired posterior predictive distribution. For example, the Bayesian equivalent to Eq. (2.11) is computed $\hat{\mu}_j = \sum_{s=1}^{M} \mu_j^{(s)}/M$.

2.5.3 Metrics for comparison

We now discuss the metrics used for evaluating estimators in Section 3.1 and unit-level predictions via cross-validation in Section 3.2. Since we evaluate both unit-level predictions and areal estimates with these metrics, we use generic θ and $\hat{\theta}$ to denote a parameter of interest and an estimate of that parameter of interest, respectively.

First the root mean square error (RMSE) is as follows,

$$\text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \theta)^2}, \tag{2.15}$$

where m is the number of estimates produced for the given parameter of interest. We evaluate this metric empirically across simulation repetitions and hold-out sets for areal estimates and unit-level predictions, respectively. Next, the bias is as follows

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta, \tag{2.16}$$

where $\mathbb{E}(\hat{\theta})$ is computed empirically across the design-based samples. We use the bias metric to evaluate the bias of estimates, estimates of the RMSE, and unit-level predictions in subsequent sections. Similar to the RMSE, the bias metric is evaluated empirically. Finally, the indicator of coverage for a given uncertainty interval C is

$$\mathbb{1}_C(\hat{\theta}) = \begin{cases} 1, & \text{if } \hat{\theta} \in C, \\ 0, & \text{if } \hat{\theta} \notin C. \end{cases} \tag{2.17}$$

2.6 Simulation

A simulation study was used to assess qualities of the estimators introduced in Section 2.2. Simulation studies are particularly useful in SAE research as they allow for us to assess estimators against simulated true parameter values, letting us gain accurate insight into estimator performance. In order to assess estimators in a manner that is fair and realistic, we utilized methodology introduced in White et al. (2024a) for our simulation study. Briefly, the approach uses a k nearest neighbors (kNN) algorithm on auxiliary data weighted with bootstrap inclusion probabilities to impute forest inventory attributes at each population unit. Formally, the algorithm to generate the population for a given state is defined in White et al. (2024a) Algorithm 1.

It is important to note that we generated simulated populations separately in Washington and Nevada. Further, as discussed in Section 2.1, we used `tcc`, `elev`, and `ppt` as auxiliary data in the kNN matching in Washington, and `tcc`, `elev`, and `tri` in Nevada. We also used `tnt` as a stratification variable for generating the simulated population. These auxiliary variables are centered to mean 0 and scaled to standard deviation 1 before the matching occurs. Without the centering and scaling step, variables of different magnitudes would hold different weight in the kNN

search. We implemented the generation of the simulated population via the `kbaabb` R package (White et al., 2024b).

After generating the simulated population we took design-based samples from the simulated population for the purposes of estimator assessment. In order to create one sample from the simulated population, we took a simple random sample from each county of the same size as the number of FIA plots in that county. Therefore, both the county and overall sample sizes remained constant.

CHAPTER 3

RESULTS

In this chapter, we discuss the results of the simulation study and FIA data application in Sections 3.1 and 3.2, respectively. In Section 3.1, to evaluate performance of the estimators, we assess metrics introduced in Section 2.5.3. The FIA data application presents county-level biomass estimates and compares estimators based on cross-validation.

3.1 Simulation study

We evaluated the nine estimators introduced in Chapter 2 across all samples taken from the simulated population generated in Section 2.6. The models used in these estimators were fit separately between the two states considered in this study, Nevada and Washington. We evaluated the performance metrics in Washington and Nevada. Generally, estimator performance differs between the two states, likely due to the substantial differences in their ecological landscapes, with Nevada's forests primarily occupying high elevation sky islands, and Washington's forests primarily on the west side of the Cascade mountain range.

Figure 3.1 displays values for the four performance metrics evaluated in Nevada, with each point on the plot representing a county-performance metric-estimator combination. In Figure 3.1a, we see that the single stage estimators exhibit the most bias. Further, we see that estimators that do not allow for county-specific variances tend to exhibit more empirical bias than those that allow for county-specific variances. The least biased estimators are the B ZI CVI CRV and the B ZI CVI SVI CRV, both performing quite similar to each other. The CVC models, when compared to their CVI counterparts, tend to exhibit more empirical bias after CRV and/or SVI terms are added. When looking at Figure 3.1b, the negative effect of the CVC effects in Nevada becomes even clearer, as we see these models exhibit much higher empirical RMSE than those without the CVC terms. All two-stage estimators with the CVI effect perform comparably in terms of RMSE, with the others performing comparably to each other. Figure 3.1c displays the empirical bias of $\widehat{\text{RMSE}}$, indicating if an estimator is over- or under-estimating its variability. The B CVI and B CVC tend to substantially under estimate their variability, likely due to the poorly specified underlying

model. Somewhat surprisingly, despite accounting for the zero-inflation in the data, the B ZI CVI estimator also exhibits some bias. The F ZI CVI does quite a good job of estimating its RMSE, but has more variability in those estimates than the B ZI CVC, B ZI CVI CRV, B ZI CVC CRV, B ZI CVI SVI CRV, and B ZI CVC SVI CRV, all of which estimate the RMSE quite well, but have a few negative outliers. Figure 3.1d displays the empirical 95% uncertainty interval coverage, where we can see the B CVI and B CVC estimators performing very poorly. The F ZI CVI performs better than its Bayesian counterpart in this case, but both have some counties with very low coverage rates. The remaining estimators all perform quite well and similar to each other, with the B ZI CVI SVI CRV performing the best on median.

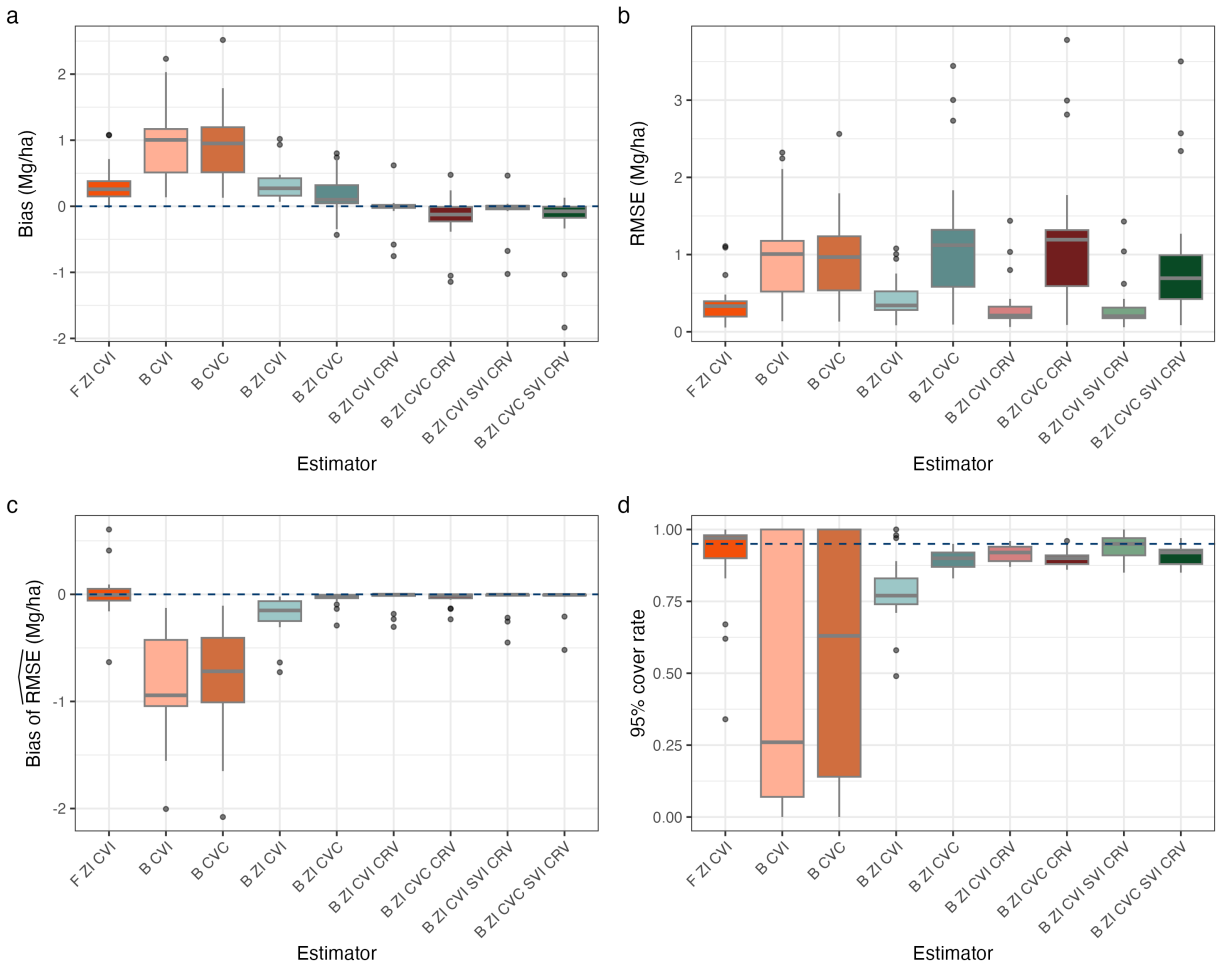

Figure 3.1 Estimator performance metrics in Nevada. The x-axis and fill corresponds to estimator and the y-axis corresponds to the value of the performance metric. Each point represents the performance metric in a given county for a particular estimator. (a) shows the estimator's empirical bias, (b) shows the estimator's empirical root mean square error (RMSE), (c) shows the empirical bias of the RMSE estimator, and (d) shows the estimator's empirical 95% uncertainty interval coverage rate. Abbreviations are: root mean square error (RMSE); frequentist (F); zero-inflated (ZI); county-varying intercept (CVI); Bayesian (B); county-varying coefficients (CVC); county-specific residual variance (CRV); space-varying intercept (SVI).

Figure 3.2 displays values for the four performance metrics evaluated in Washington, with each point on the plot representing a county-performance metric-estimator combination. Figure 3.2a displays the empirical bias of each estimator, and we see in Washington that the F ZI CVI estimator exhibits the most bias, substantially more than its Bayesian counterpart. Initially this may come as a surprise due to the estimators similar structure, but one must consider the differences in the back-

transformation of the response variable between these estimators. While the frequentist estimator back-transforms the predictions from each model, the Bayesian estimators back-transforms each sample from the posterior predictive distribution, potentially leading to more sensible estimates under certain conditions. The B CVI and B CVC estimators tend to exhibit greater magnitude bias than the two-stage Bayesian estimators, all of which exhibit similar amounts of bias to each other. One notable observation is the B ZI CVC CRV and B ZI CVC SVI CRV estimators exhibit a slight amount more empirical bias than their CVI counterparts; and in Figure 3.2b we can see their empirical RMSE is higher than their CVI counterparts. The B ZI CVI CRV and B ZI CVI SVI CRV estimators have the lowest empirical RMSE, while the F ZI CVI estimator has the highest empirical RMSE. Other two-stage Bayesian estimators exhibit very similar empirical RMSE to each other, and the single-stage estimators exhibit a bit higher empirical RMSE. Turning to Figure 3.2c, we see the empirical bias of the RMSE estimator. Notably, the F ZI CVI and B CVI estimators exhibit the most bias here, while the other estimators exhibit very low bias. Once CRVs are accounted for, we see almost no bias. Finally, turning to Figure 3.2d, we examine the empirical 95% uncertainty interval coverage rate. The B CVI and B CVC estimators have very poor coverage, while the other estimators, all of which are two-stage, have decent coverage among themselves. Notably, the estimators which account for CRVs and have CVC effects exhibit the best coverage rates.

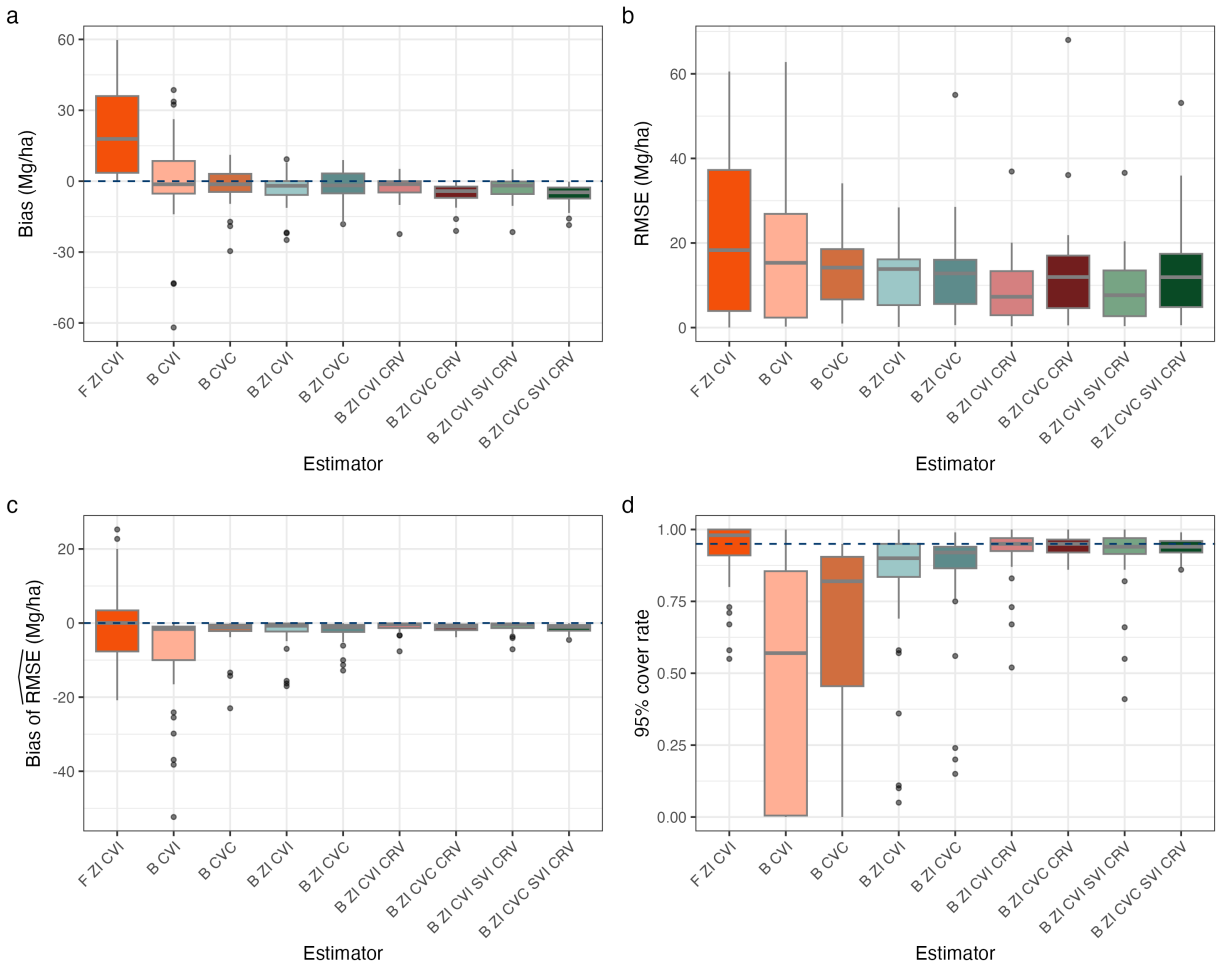

Figure 3.2 Estimator performance metrics in Washington. The x-axis and fill corresponds to estimator and the y-axis corresponds to the value of the performance metric. Each point represents the performance metric in a given county for a particular estimator. (a) shows the estimator's empirical bias, (b) shows the estimator's empirical root mean square error (RMSE), (c) shows the empirical bias of the RMSE estimator, and (d) shows the estimator's empirical 95% uncertainty interval coverage rate. Abbreviations are: root mean square error (RMSE); frequentist (F); zero-inflated (ZI); county-varying intercept (CVI); Bayesian (B); county-varying coefficients (CVC); county-specific residual variance (CRV); space-varying intercept (SVI).

Stepping back from the details of Figures 3.1 and 3.2, we can gain some broader insights about the estimators and the simulation study. First of all, it is striking that the estimators that include a SVI perform almost identically to their non-spatial counterparts. This is not to say that spatial effects are not useful for estimating biomass or other forest attributes, in fact we know from other studies these effects are quite useful (Finley et al., 2024). Further, from the cross-validation carried out in

Section 3.2 we see more substantial improvements with the spatial models. It is the case though, that the simulated population we generated does not necessarily reflect the spatial structure of the true population, and in fact upon inspection it has very little spatial structure. Incorporating some spatial smoothing into the population generation methodology would help improve the simulated population's utility for assessing these spatial models. Another broad observation we can make from the simulation study is that the CVCs generally did not improve estimation, and in some cases they worsened estimation. We see this particularly in Nevada, where county-level biomass is more homogeneous across the state. In Washington, the estimators with CVCs did have better 95% coverage rates than their CVI counterparts, but at the cost of higher bias and RMSE. The use of CVCs may be more useful for models fit across larger spatial domains (such as the entire United States), where we would expect the effect of predictors on the response to vary substantially. States with more diverse forest types may also benefit from the use of CVCs.

3.2 FIA data application

We applied the estimators discussed to actual plot data collected by FIA in both Nevada and Washington. We first compare the estimators using 10-fold cross-validation, and then turn to discussing estimates produced by the B ZI CVI SVI CRV estimator, an estimator that showed favorability in the simulation study and in cross-validation.

We performed 10-fold cross-validation at the unit-level to help assess the estimators' performance, with the unit-level predictions as a proxy for estimation of county means. Results from the cross-validation are shown in Table 3.1. To compute 95% uncertainty interval coverage rates ("Coverage" in Table 3.1), we used the quantiles of the posterior predictive distribution of $y(\ell)$ for the Bayesian estimators. We omit 95% confidence interval coverage for the frequentist estimators as the task of computing closed form or bootstrap prediction intervals at the unit-level for this two-stage model is beyond the scope of this article. Regarding the results, we can see similar patterns as we saw in Figures 3.1 and 3.2, but now we can better see the performance improvement associated with the spatial models when considering the root mean square prediction error (RMSPE) and bias. We see that across states, empirical bias and coverage rates were favorable for all estimators

that included a county-specific variance term, and in Nevada for all estimators that account for zero-inflation. Examining the results of this cross-validation is helpful for model evaluation and in this case provides similar insights to the simulation study, and in Chapter 4 we discuss the tradeoffs of the two model evaluation methods.

State	Metric	F ZI CVI	B CVI	B CVC	B ZI CVI	B ZI CVC	B ZI CVI CRV
WA	RMSPE	107.05	113.90	111.48	106.87	104.78	104.03
WA	Bias	18.10	22.49	19.75	17.52	15.49	-1.63
WA	Coverage	NA	42.95	49.10	65.00	69.41	96.77
NV	RMSPE	7.79	8.17	8.27	7.79	9.30	7.82
NV	Bias	0.35	0.85	0.88	0.33	-0.24	0.00
NV	Coverage	NA	49.95	68.79	91.57	95.41	97.85

State	Metric	B ZI CVC CRV	B ZI CVI SVI CRV	B ZI CVC SVI CRV
WA	RMSPE	103.17	98.62	99.15
WA	Bias	-3.21	-2.15	-3.71
WA	Coverage	96.59	96.71	96.68
NV	RMSPE	8.90	7.77	7.73
NV	Bias	-0.35	0.00	-0.02
NV	Coverage	97.99	98.04	97.92

Table 3.1 Results for each estimator and state from the unit-level cross-validation. Results include empirical measures of root mean square prediction error, bias, and 95% uncertainty interval coverage rate. Abbreviations are: Washington (WA); Nevada (NV); root mean square prediction error (RMSPE); frequentist (F); zero-inflated (ZI); county-varying intercept (CVI); Bayesian (B); county-varying coefficients (CVC); county-specific residual variance (CRV); space-varying intercept (SVI).

In both Nevada and Washington, the B ZI CVI SVI CRV estimator performs quite well when considering all computed metrics in the simulation study and cross-validation, so we chose to use this estimator as an example for producing pixel-level predictions and county-level estimates in Washington and Nevada. Figure 3.3a, b show pixel-level estimates of biomass probability and Figure 3.3c, d show estimated biomass (Mg/ha) in both Washington and Nevada produced by the B ZI CVI SVI CRV estimator. While we are not particularly interested in the pixel-level predictions for the study at hand, these sorts of maps can be very useful for management at very small spatial scales, such as the stand level. In this Bayesian framework, not only do we have pixel-level predictions, but we also have pixel-level uncertainty predictions. Further, we can easily summarize these pixel-level predictions to get biomass estimates and uncertainty estimates for any area that

may be of interest. For the study at hand, we only produce estimates for counties, but we could have just as easily produced estimates for any custom small area of interest (e.g., ecoregions, watersheds, fires, stands, etc.) with any of the Bayesian estimators.

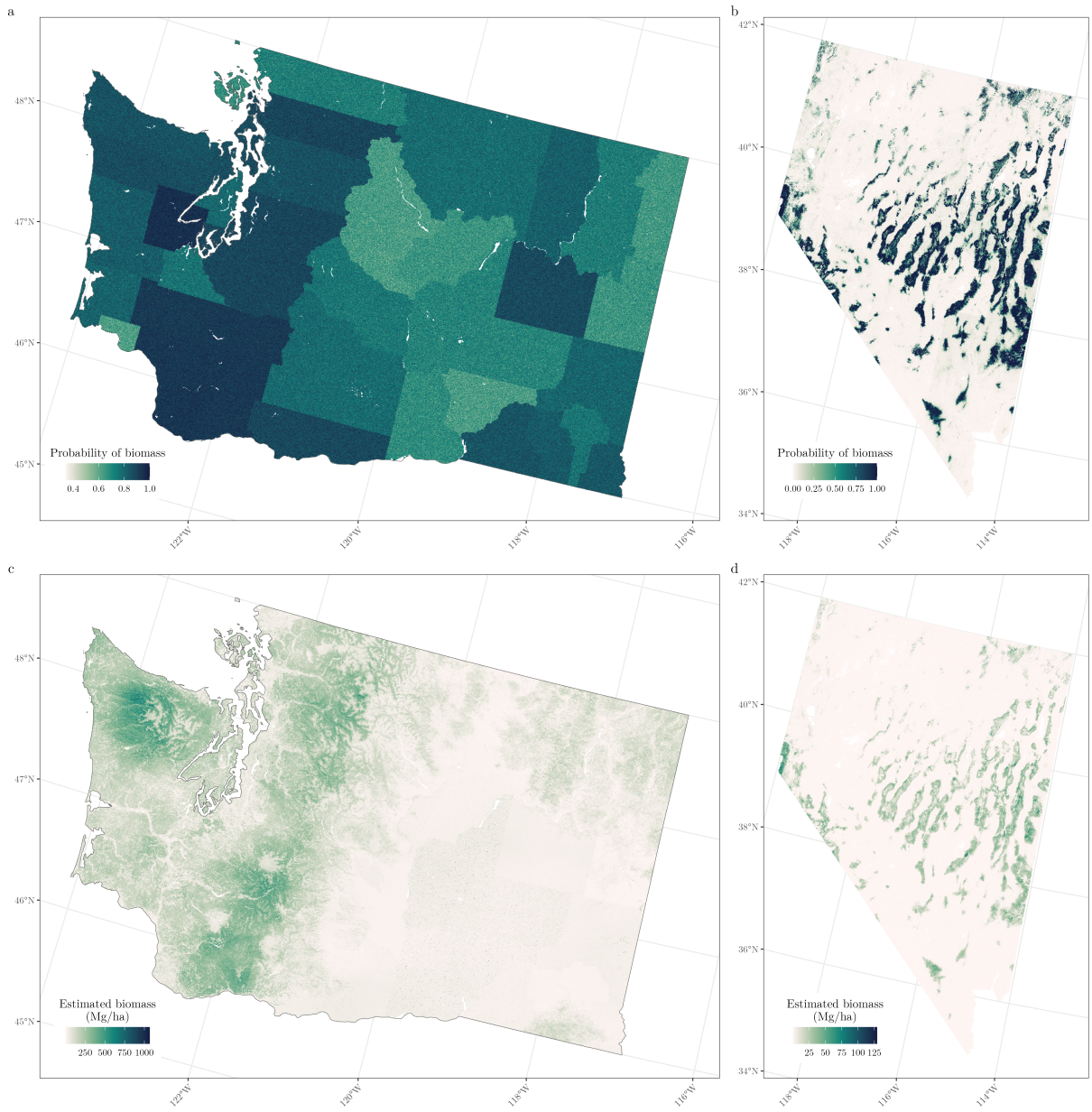

Figure 3.3 Pixel-level estimates of biomass probability (a, b) and estimated biomass (c, d) in Washington (a, c) and Nevada (b, d). Pixel-level estimates of biomass probability are produced from the Bayesian model with Bernoulli response and county varying intercept. Pixel-level estimates of biomass are produced from the Bayesian zero-inflated model with county varying intercept with space varying intercept and county specific variances.

Figure 3.4 shows the estimated county-level biomass (Mg/ha) in both Nevada and Washington produced by the B ZI CVI SVI CRV estimator. These sorts of estimates are what we are most interested for SAE, and for demonstration we only produce estimates at the county-level. County-level estimates of biomass show the stark change in biomass across the cascades in Washington state, and the relatively constant county-level biomass across the state of Nevada.

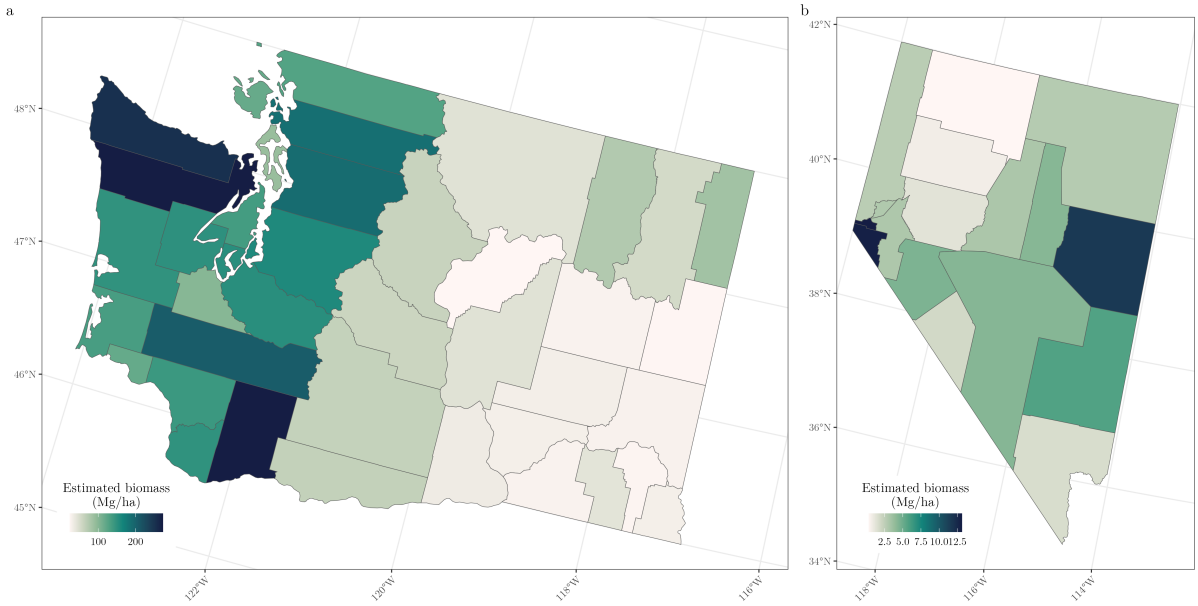

Figure 3.4 County-level estimates of average biomass (Mg/ha) in Washington (a) and Nevada (b). Estimates are produced from the Bayesian zero-inflated model with county varying intercept with space varying intercept and county specific variances.

CHAPTER 4

DISCUSSION

Our work implemented and compared nine model-based approaches from both the frequentist and Bayesian statistical paradigms to estimating biomass in counties in the states of Nevada and Washington. The Bayesian estimators used allow for flexibility in model specification, allowing for CVIs, CVCs, CRVs, and SVIs. We assessed the proposed estimators through a simulation study using design-based samples from a simulated population generated by a bootstrap-weighted kNN technique, and through cross-validation at the unit-level on FIA data. Results suggest that accounting for zero inflation through a two-stage approach, the inclusion of CRVs, and the inclusion of SVIs yield the best performing estimators.

Our model-based approaches to estimation of biomass in these zero-inflated small areas can be classified as a *unit-level* approach to small area estimation, meaning that the models used were fit to data at the survey unit level. By fitting our models at the unit-level, we had to account for zero-inflation by utilizing a two-stage modeling approach. Another commonly used small area approach for this sort of problem is to fit an *area-level* model, where the model is fit to data aggregated to the area of interest. When the data are aggregated to the area-level, the issue of zero-inflation ceases to persist, as long as some positive amount of biomass has been sampled in each area of interest. These area-level models are attractive due to their simplicity, interpretability, and ease of implementation. Cao et al. (2022) fit spatial and non-spatial area-level models for small area estimation of county-level timber volume and found increased precision over a direct, design-based approach. However, in White et al. (2025) the F ZI CVI estimator showed increased precision and reduced bias over an area-level approach, which lead us to take unit-level zero-inflated approaches in this study. There are other approaches to dealing with zero-inflation in the unit-level data that we did not consider in this study such as non-linear models and non-parametric approaches. For example, McRoberts et al. (2007) suggest a kNN approach for estimation of forest attributes, but in our study we only considered models with higher amounts of interpretability than kNN. A more interpretable non-parametric approach to consider would be regression trees. McConville and Toth

(2019) present a model-assisted method for automating post-stratification via regression trees and apply their method to labor and employment survey data, but this approach may be useful in our forest inventory setting.

A curious result from our study resides in the F ZI CVI estimator performing substantially differently than the B ZI CVI estimator in Washington. Initially, we would expect these estimators to perform almost identically given the model forms used in estimation, and this is what occurs in Nevada. The difference between these estimators that causes such a large discrepancy in estimates lies within the differences in how we back-transform the response variable in the frequentist and Bayesian settings. For the frequentist approach, we back-transform the response using the maximum likelihood point estimate at some location $\hat{y}(\ell)$, but in the Bayesian approach we back-transform each sample from the posterior predictive distribution of the response at that location ℓ. In Washington, biomass is more heavily skewed than in Nevada and the back-transformation of only the point estimate does not seem to produce a realistic prediction in these highly skewed locations, resulting in bias in the estimates. However, with the Bayesian approach, we are able to preserve the non-symmetric distribution of biomass at that location, leading to more realistic estimates.

Comparing estimators for the purposes of SAE is a difficult problem given the small sample size in areas of interest and lack of true parameter values at the granularity of the area of interest. We implemented two methods for evaluating the proposed estimators, each of which have tradeoffs. First, we generated a simulated population for estimation evaluation. This approach allows for us to assess estimators empirically by sampling repeatedly from the simulated population and comparing estimates to the true value for the response in each county. These assessments are only insightful in the case that (1) the simulated population does not unfairly portray the estimators by being generated from some method too similar to the proposed estimators; and (2) the simulated population is similar enough to the true population that fitting these estimators to samples from the simulated population is representative of fitting the estimators to a sample of the true population. We achieve (1) here by using the bootstrap-weighted kNN simulated population generation technique proposed by White et al. (2024a), but partially miss the mark regarding (2) as the simulated population does

not have similar spatial structure to the true population, impeding our ability to accurately assess the models with a space varying intercept. The bootstrap-weighted kNN approach produced a simulated population with very little spatial correlation of biomass, which is not how we expect the true population to be structured. This artifact lead the models with space varying intercepts to not improve estimation as much as we would expect them to. The benefit of the models with space varying intercepts could be clearly seen in the cross-validation that we used to assess predictions made onto unit-level observations. This approach does not require assumptions about how the population is generated or otherwise, but only allows for assessment of unit-level predictions. In our case, the cross-validation analysis yielded very similar results to the simulation study, and we were able to more clearly see the benefit of the spatial models. However, cross-validation is only possible for unit-level small area estimators and we must tolerate unit-level predictions as a proxy for small area estimates.

Our analyses provide insight as to how we may want to go about estimating biomass, or other continuous zero-inflated forest attributes, in areas of mixed landscape types and how a practitioner may want to assess their modeling choices. Given our changing climate and increasingly prolific disturbances, such as fires, producing accurate estimates of forest attributes in small areas that are zero-inflated is of the upmost importance for understanding our ever-changing landscape. A limitation of this study is the lack of spatial structure preserved by the methodology for creating simulated populations. In order to better assess the quality of spatial area- and unit-level small area estimators, we plan to add spatial structure to our simulated population method via non-parametric spatial smoothing approaches. In future work, we hope to test the proposed estimators across a variety of regions such as recently burned areas, forest stands, and other areas with great ecological importance. We also hope to consider temporal effects in these zero-inflated estimators, to be able to estimate change in biomass across time scales.

BIBLIOGRAPHY

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Burrill, E., Christensen, G., Conkling, B., DiTommaso, A., Lepine, L., Perry, C., Pugh, S., Turner, J., Walker, D., and Williams, M. (2023). Forest inventory and analysis database: Database description and user guide for phase 2 (version: 9.1).

Cao, Q., Dettmann, G. T., Radtke, P. J., Coulston, J. W., Derwin, J., Thomas, V. A., Burkhart, H. E., and Wynne, R. H. (2022). Increased precision in county-level volume estimates in the united states national forest inventory with area-level small area estimation. *Frontiers in Forests and Global Change*, 5:769917.

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

Daly, C., Gibson, W. P., Taylor, G. H., Johnson, G. L., and Pasteris, P. (2002). A knowledge-based approach to the statistical mapping of climate. *Climate Research*, 22(2):99–113.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

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

Finley, A. O. (2025). *Models for zero-inflated data*. Available at `https://github.com/finleya/zi_models`.

Finley, A. O., Andersen, H.-E., Babcock, C., Cook, B. D., Morton, D. C., and Banerjee, S. (2024). Models to support forest inventory and small area estimation using sparsely sampled lidar: A case study involving g-liht lidar in tanana, alaska. *Journal of Agricultural, Biological and Environmental Statistics*, 29(4):695–722.

Finley, A. O., Banerjee, S., and MacFarlane, D. W. (2011). A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. *Journal of the American Statistical Association*, 106(493):31–48.

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2):401–414.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

May, P., McConville, K. S., Moisen, G. G., Bruening, J., and Dubayah, R. (2023). A spatially varying model for small area estimates of biomass density across the contiguous united states. *Remote Sensing of Environment*, 286:113420.

McConville, K. S. and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2):389–413.

McRoberts, R. E., Tomppo, E. O., Finley, A. O., and Heikkinen, J. (2007). Estimating areal means and variances of forest attributes using the k-nearest neighbors technique and satellite imagery. *Remote sensing of environment*, 111(4):466–480.

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

Picotte, J. J., Dockter, D., Long, J., Tolk, B., Davidson, A., and Peterson, B. (2019). LANDFIRE remap prototype mapping effort: Developing a new framework for mapping vegetation classification, change, and structure. *Fire*, 2(2):35.

Prisley, S., Bradley, J., Clutter, M., Friedman, S., Kempka, D., Rakestraw, J., and Sonne Hall, E. (2021). Needs for small area estimation: Perspectives from the us private forest sector. *Frontiers in Forests and Global Change*, 4:746439.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, J. and Molina, I. (2015). *Small Area Estimation*. Wiley, 2nd edition. ISBN: 978-1-118-73578-7.

Rollins, M. G. (2009). LANDFIRE: a nationally consistent vegetation, wildland fire, and fuel assessment. *International Journal of Wildland Fire*, 18(3):235–249.

U.S. Geological Survey (2019). LANDFIRE Elevation.

U.S. Senate (2023). *S. Rep. 118-83 - Department of the Interior, Environment, and Related Agencies Appropriations Bill*.

White, G. W., Wieczorek, J. A., Cody, Z. W., Tan, E. X., Chistolini, J. O., McConville, K. S., Frescino, T. S., and Moisen, G. G. (2024a). Assessing small area estimates via bootstrap-weighted k-nearest-neighbor artificial populations.

White, G. W., Wieczorek, J. A., Frescino, T. S., and McConville, K. S. (2024b). *kbaabb: Generates*

an Artificial Population Based on the KBAABB Methodology. R package version 0.0.0.9000.

White, G. W., Yamamoto, J. K., Elsyad, D. H., Schmitt, J. F., Korsgaard, N. H., Hu, J. K., Gaines, G. C., Frescino, T. S., and McConville, K. S. (2025). Small area estimation of forest biomass via a two-stage model for continuous zero-inflated data. *Canadian Journal of Forest Research*, 55:1–19.

Wiener, S. S., Bush, R., Nathanson, A., Pelz, K., Palmer, M., Alexander, M. L., Anderson, D., Treasure, E., Baggs, J., and Sheffield, R. (2021). United states forest service use of forest inventory data: Examples and needs for small area estimation. *Frontiers in Forests and Global Change*, 4:763487.

Yamamoto, J., Elsyad, D., White, G., Schmitt, J., Korsgaard, N., McConville, K., and Hu, K. (2025). *saeczi: Small Area Estimation for Continuous Zero Inflated Data*. R package version 0.2.0.9000, commit f580319049b152c890033c770913b7a296ce63cb.

Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S. M., Case, A., Costello, C., Dewitz, J., Fry, J., Funk, M., Granneman, B., Liknes, G. C., Rigge, M., and Xian, G. (2018). A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:108–123.

APPENDIX

A. PRIOR DISTRIBUTIONS AND HYPERPARAMETERS

Table A1 includes the prior distribution and hyperparameter values for all parameters used for the Bayesian models.

Parameter	Prior distribution	hyperparameter values
β_0	$\mathcal{N}(\mu,\ \sigma^2)$	$\mu = 0,\ \sigma^2 = 1000$
$\tilde{\beta}_0(\ell)$	$\mathcal{N}(\mu,\ \sigma^2)$	$\mu = 0,\ \sigma^2 = \sigma^2_{\tilde{\beta}_0}$
β_k	$\mathcal{N}(\mu,\ \sigma^2)$	$\mu = 0,\ \sigma^2 = 1000$
$\tilde{\beta}_{k,j}$	$\mathcal{N}(\mu,\ \sigma^2)$	$\mu = 0,\ \sigma^2 = \sigma^2_{\tilde{\beta}_{k,j}}$
$\sigma^2_{\tilde{\beta}_0}$	$\mathcal{IG}(a,\ b)$	$a = 2,\ b = 1$
$\sigma^2_{\tilde{\beta}_{k,j}}$	$\mathcal{IG}(a,\ b)$	$a = 2,\ b = 1$
$\varepsilon(\ell)$	$\mathcal{N}(\mu,\ \sigma^2)$	$\mu = 0,\ \sigma^2 = \tau^2$
τ^2	$\mathcal{IG}(a,\ b)$	$a = 2,\ b = 10$
α_0	$\mathcal{N}(\mu,\ \sigma^2)$	$\mu = 0,\ \sigma^2 = 1000$
$\tilde{\alpha}_0(\ell)$	$\mathcal{N}(\mu,\ \sigma^2)$	$\mu = 0,\ \sigma^2 = \sigma^2_{\tilde{\alpha}_0}$
α_k	$\mathcal{N}(\mu,\ \sigma^2)$	$\mu = 0,\ \sigma^2 = \sigma^2_{\alpha_k}$
$\sigma^2_{\tilde{\alpha}_0}$	$\mathcal{IG}(a,\ b)$	$a = 2,\ b = 1$
$\boldsymbol{w}$	$\mathcal{MVN}(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$	$\boldsymbol{\mu} = \mathbf{0},\ \boldsymbol{\Sigma} = \sigma^2_w \boldsymbol{R}(\phi)$
σ^2_w	$\mathcal{IG}(a,\ b)$	$a = 2,\ b = 1$
ϕ	$\mathcal{U}(a,\ b)$	$a = 0.003,\ b = 3$
$\varepsilon(\ell)$	$\mathcal{N}(\mu,\ \sigma^2)$	$\mu = 0,\ \sigma^2 = \tau^2$
$\varepsilon_1(\ell)$	$\mathcal{N}(\mu,\ \sigma^2)$	$\mu = 0,\ \sigma^2 = \tau^2_1$ or $\tau^2_{1,j}$
$\varepsilon_2(\ell)$	$\mathcal{N}(\mu,\ \sigma^2)$	$\mu = 0,\ \sigma^2 = \tau^2_2$
τ^2	$\mathcal{IG}(a,\ b)$	$a = 2,\ b = 10$
τ^2_1	$\mathcal{IG}(a,\ b)$	$a = 2,\ b = 10$
$\tau^2_{1,j}$	$\mathcal{IG}(a,\ b)$	$a = 2,\ b = 10$
τ^2_2	None	Set to 0.00001

Table A1 Prior distributions and hyperparameter values used for the Bayesian models.