A Spatio-Temporal Model for Mountain Pine Beetle Damage

Kimberly A. KAUFELD, Matthew J. HEATON, and Stephan R. SAIN

Forest composition in the western region of the United States has seen a dramatic change over the past few years due to an increase in mountain pine beetle damage. In order to mitigate the pine beetle epidemic, statistical modeling is needed to predict both the occurrence and the extent of pine beetle damage. Using data on the front range mountains in Colorado between the years 2001–2010 from the National Forest Service, we develop a zero-augmented spatio-temporal beta regression model to predict both the occurrence of pine beetle damage (a binary outcome) and, given damage occurred, the percent of the region infected. Temporal evolution of the pine beetle damage is captured using a dynamic linear model where both the probability and extent of damage depend on the amount of damage incurred in neighboring regions in the previous time period. A sparse conditional autoregressive model is used to capture any spatial information not modeled by spatially varying covariates. We find that the occurrence and extent of pine beetle damage are positively associated with slope and damage in previous time periods.

Key Words: Zero-augmented; Beta regression; Stick-breaking; Dimension reduction.

1. INTRODUCTION

1.1. Problem Statement and Description of Data

The mountain pine beetle *Dendroctonus ponderosae* (MPB) is an insect that burrows, resides, and reproduces in mature pine stands. Native to the forests of the western United States, MPBs have, historically, played an important role in forest health by attacking weakened trees—thus speeding development of a younger, more healthy forest. However, the recent onset of warm summers and dry conditions has created an epidemic (Williams and Liebhold 2002). In particular, multiple MPB outbreaks have caused wide spread tree mortal-

Kimberly A. Kaufeld (✉) is Postdoctoral Researcher in Statistical and Applied Mathematical Sciences Institute, 19 T.W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC 27709, USA (E-mail: *kimberly.kaufeld@gmail.com*). Matthew J. Heaton is Assistant Professor in Department of Statistics, Brigham Young University, 204 TMCB, Provo, UT 84602, USA (E-mail: *mheaton@stat.byu.edu*). Stephan R. Sain is Scientist in National Center for Atmospheric Research, Institute for Mathematics Applied to Geosciences, P.O. Box 3000, Boulder, CO 80307, USA (E-mail: *ssain@ucar.edu*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 19, Number 4, Pages 439–452
DOI: 10.1007/s13253-014-0182-1

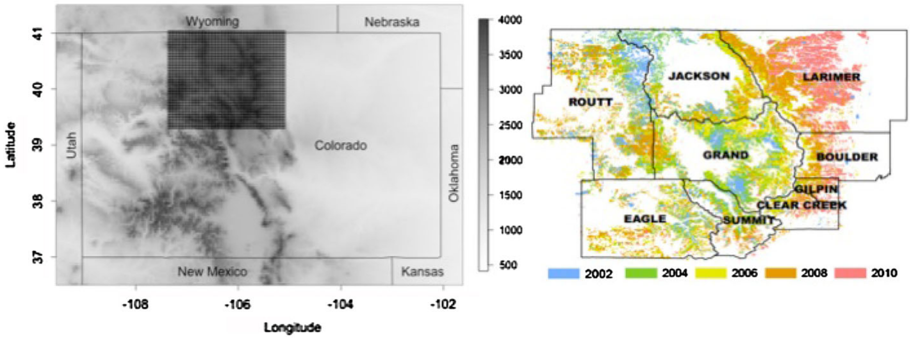

Figure 1. (*left*) Elevation map of Colorado with overlaid spatial grid of study region and (*right*) example of Colorado counties impacted by MPB damage contained within the spatial grid of the study region from 2002 to 2010.

ity in conifer forests including ponderosa and lodgepole pines since the early 1990s (Raffa et al. 2008).

To estimate the extent of MPB damage over a region, ecologists often rely on annual aerial detection surveys (ADS) to analyze spatial and temporal patterns of the damage (Harris et al. 2002, 2003). The primary motivation for this article comes from one such ADS conducted by the Colorado State Forest Service (CSFS) for the front range mountains in Colorado during the years 2001–2010. Each year, surveyors would fly over the survey area and digitally draw in regions on a map to denote damaged areas. The ADS extends from the southern Rocky Mountains in Colorado to southern Wyoming and the Black Hills of South Dakota, but we focus on analyzing the data in the North Central Rocky Mountains in Colorado because this area has more consistent pine tree cover. The gridded area in the left panel of Fig. 1 displays the region of interest for this study.

When considered temporally, the ADS data represent a cumulative summary of damaged areas. Notationally, let $\mathscr{R}_{it}$ be the ith damaged region drawn in year t. For example, $\mathscr{R}_{it}$ represents one of the highlighted areas in the right panel of Fig. 1. The damaged areas in year t are, then, the union of all the damaged regions drawn in years up to and including time t. Mathematically, damaged regions in year t ($\mathscr{D}_t$) are given by $\mathscr{D}_t = \cup_{t' \leq t} \cup_{i=1}^{R_{t'}} \mathscr{R}_{it'}$ where R_t are the total number of damaged regions drawn in year t.

Statistically summarizing and modeling ADS data are an interesting challenge that can benefit researchers by helping to make informed decisions for ground management actions and aerial surveying based upon the probable damage in a particular area. Zhu et al. (2005, 2008); Zheng and Zhu (2008) consider data aggregated to a regular lattice where each grid cell is assigned a binary response (infected or not infected) and use autologistic models to model the spatio-temporal structure. We note that this type of aggregation for our ADS data, however, would result in information loss. That is, classifying a grid cell as "infected" acts as if the whole region has been infected when in reality only a portion of the region may have been impacted. Proportions of damaged areas better capture the nature of the ADS data than binary models.

To make ADS data more amenable to statistical modeling and following previous studies of MPB damage, we aggregated the ADS data to a spatial grid with 42 rows and 55 columns

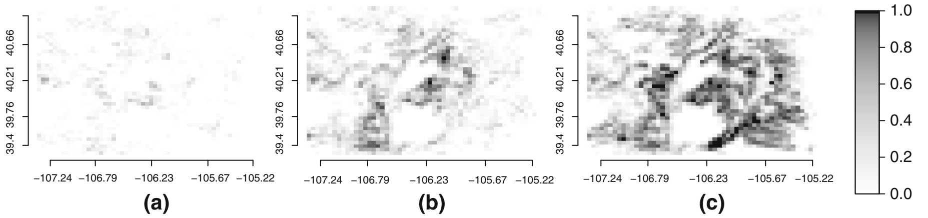

Figure 2. Observed $\widetilde{y}_{gt}$ for **a** $t = 2001$, **b** $t = 2005$ and **c** $t = 2010$. Note the strong spatial correlation between sites and the monotonically increasing proportion of damaged area.

($G = 2310$ total grid cells). On this grid, each cell represents an area of roughly 16 km^2. This grid size was chosen for three reasons. First, a resolution of 4 km is fine enough to capture landscape variability in addition to climate variation between cells. Second, this aligns with resolution of the meteorological dataset from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) (Daly et al. 2002) used in the analysis below. The left panel of Fig. 1 displays the spatial grid.

In order to minimize information loss due to aggregation, the cumulative damage (rather than a binary summary) for year $t = 1, \ldots, 10$ where $t = 1$ corresponds to the year 2001 was calculated for each grid cell. That is, for each year, we calculated the percent of a grid cell that fell within a damaged region in that year or any previous year. More concretely, let $\widetilde{y}_{gt}$ represent the cumulative MPB damage in grid cell g up to year t. The $\widetilde{y}_{gt}$ are calculated as

$$\widetilde{y}_{gt} = \frac{1}{|\mathscr{G}_g|} \int_{\mathscr{G}_g} \mathbb{1}_{\{s \in \mathscr{D}_t\}} ds \tag{1.1}$$

where $\mathscr{G}_g$ represents the spatial region of grid cell g, $\mathbb{1}_{\{\cdot\}}$ is an indicator function and $\mathscr{D}_t = \cup_{t' \leq t} \cup_{i=1}^{R_{t'}} \mathscr{R}_{it'}$ are the damaged regions up to year t. Some of the observed $\widetilde{y}_{gt}$ are displayed in Fig. 2. Note that $\widetilde{y}_{gt} \in [0, 1)$ where $\widetilde{y}_{gt} \neq 1$ because grid cells never reach a "completely damaged" state. Furthermore, the $\widetilde{y}_{gt}$ are monotonically increasing as a function of t because, for the ADS data, the region of damaged areas only increases (once damaged, a grid cell will always be damaged).

1.2. ARTICLE OVERVIEW AND OUTLINE

The goal of this work is to develop a modeling strategy for $\widetilde{y}_{gt}$ to aid in understanding and predicting MPB damage. The ultimate goal is to develop intervention strategies to prevent further damage. Because the $\widetilde{y}_{gt} \in [0, 1)$ and are monotonic, beta regression models advocated by Kieschnick and McCullough (2003) and Ferrari and Cribari-Neto (2004) are not a viable modeling strategy as these are only defined on the open interval (0, 1). More appropriate is the work by Ospina and Ferrari (2010); Wieczorek and Hawala (2011); Ospina and Ferrari (2012), and Wieczorek et al. (2012) who develop zero, one and zero-and-one-augmented beta regression models. Perhaps most pertinent to the data described here is the work by Hatfield et al. (2012) who develop a zero-augmented beta regression model with individual-specific latent trajectories to explain the probability of a zero outcome and the mean of a non-zero outcome. None of these approaches for modeling random variables on $[0, 1)$, however, account for the monotonicity constraints which need to be imposed on the $\widetilde{y}_{gt}$.

For this article, we develop a model to explain and predict both the occurrence of pine beetle damage (a binary outcome) and, given damage occurred, the percent of the region infected. We use a stick-breaking representation to account for the monotonicity constraints of the cumulative damage ($\widetilde{y}_{gt}$) over time. Specifically, using the stick-breaking representation, the $\widetilde{y}_{gt}$ are expressed in terms of non-monotonic random variables (say, y_{gt}) with support on $[0, 1)$ and a zero-augmented spatio-temporal beta regression model is used to model y_{gt}. Following Hatfield et al. (2012), our model uses a beta regression model for proportions on (0,1) and a binary component to model the probability of no MPB damage. Our contribution, beyond the stick-breaking representation, is to add a spatial and temporal term to the model to account for the spatial and temporal variation that occurs over the Colorado region so as to exploit correlations to aid in predictions.

Temporal evolution of the pine beetle damage is captured using a dynamic linear model where both the probability and extent of damage depend on the percent of damage incurred in neighboring regions in the previous time period. The low rank conditional autoregressive (CAR) models of Hughes and Haran (2013) are used to capture any spatial information not modeled by spatially varying covariates (e.g., slope, elevation, etc.).

In Sect. 2, we use a stick-breaking representation to model the cumulative damages to enforce monotonicity, discuss the prior assumptions made for each parameter as well as outline how to perform statistical inference and prediction. Section 3 shows results of fitting the model to the MPB dataset while Sect. 4 concludes and discusses opportunities for new statistical and applied research.

2. A ZERO-AUGMENTED SPATIO-TEMPORAL MODEL FOR MOUNTAIN PINE BEETLE DAMAGE

2.1. Statistical Model

Let $g = 1, \ldots, G = 2310$ denote the grid cells of the 42×55 spatial lattice shown in the left panel of Fig. 1 and let $t = 1, \ldots, 10$ denote the year where $t = 1$ refers to the year 2001. Let $\boldsymbol{x}'_{gt} = (x_{gt1}, \ldots, x_{gtP})$ denote a vector of P covariates (e.g., elevation and precipitation; see Sect. 3). To ensure monotonicity of the cumulative damage to a grid cell, let

$$\widetilde{y}_{gt} = \sum_{t' \leq t} \left[y_{gt'} \prod_{\{i: i < t'\}} (1 - y_{gi}) \right] \tag{2.1}$$

where $y_{gt} \in [0, 1)$ are non-monotonic. The representation of $\widetilde{y}_{gt}$ in (2.1) follows the stick-breaking representation of the Dirichlet process by Sethuraman (1994). Intuitively, the y_{gt} represent the amount of MPB damage at time t to the *undamaged* portion of grid cell g. For example, at time $t = 1$, $100 \times y_{g1}$ % of the grid cell is damaged leaving $100 \times (1 - y_{g1})$ %

undamaged. At time $t = 2$, $100 \times y_{g2}$ % of the undamaged portion, $(1 - y_{g1})$, of the grid cell is damaged with a cumulative damage of $\widetilde{y}_{g2} = y_{g1} + y_{g2}(1 - y_{g1})$ and $100 \times (1 - y_{g1}) \times (1 - y_{g2})$ % undamaged. We emphasize that $y_{gt} = 0$ implies that there was no further damage at time t.

Notice that in (2.1) there is a one-to-one relationship between $\widetilde{y}_{gt}$ and y_{gt}. Hence, under the stick-breaking representation, a model $\widetilde{y}_{gt}$ is induced by modeling y_{gt}. We assume, $y_{gt} = (1 - z_{gt})b_{gt}$ where $z_{gt} \in \{0, 1\}$ is a Bernoulli random variable with $\mathbb{P}\mathrm{r}(z_{gt} = 1) = \delta_{gt} \in (0, 1)$ and $b_{gt} \in (0, 1)$. Intuitively, z_{gt} is an indicator variable for no damage, δ_{gt} is the probability that there was no damage, and b_{gt} is the amount of damage at time t conditional on the event that there was damage ($z_{gt} = 0$).

To model the probability of no damage, we assume,

$$\operatorname{logit}(\delta_{gt}) = \alpha_\delta + \boldsymbol{x}'_{gt}\boldsymbol{\beta}_\delta + \eta_{\delta g} + \phi_\delta \sum_{g' \in \mathscr{N}_g} d^{\phi_\delta}_{g'(t-1)} + \theta_\delta d^{\theta_\delta}_{g(t-1)} \tag{2.2}$$

where $\operatorname{logit}(\delta_{gt}) = \log(\delta_{gt}/(1 - \delta_{gt}))$, α_δ is an intercept, $\boldsymbol{\beta}_\delta$ is a vector of coefficients associated with $\boldsymbol{x}_{gt}$, $\eta_{\delta g}$ is a spatially correlated random effect for grid cell g designed to capture the effect of unmeasured, spatially correlated covariates associated with grid cell g, ϕ_δ is the temporal effect of damage to the neighbors ($\mathscr{N}_g$) of grid cell g, and θ_δ is the temporal effect of damage to grid cell g.

In specifying a model, we use the $d^{\phi_\delta}_{gt}$ and $d^{\theta_\delta}_{gt}$ as measures of damage to grid cell g at time t and allow them to take a value of either y_{gt} or $\widetilde{y}_{gt}$. Which measure of damage (y or $\widetilde{y}$) to use in (2.2) to capture temporal dynamics of MPB damage is not entirely clear. On one hand, it may be the case that defining $d^{\phi_\delta}_{gt} = \widetilde{y}_{gt}$ is more appropriate because MPBs will tend to migrate to a neighboring grid cell only after consuming the resources within that grid cell. On the other hand, defining $d^{\phi_\delta}_{gt} = y_{gt}$ may be more appropriate because a large value of y_{gt} could indicate a high MPB population which is likely to spread to neighboring grid cells. Using both, however, is inappropriate because there is a one-to-one correspondence between y and $\widetilde{y}$. We explore which measure of damage to use in Sect. 3.1 using variable selection.

For non-zero damage, we assume $b_{gt} \sim \mathscr{B}(\mu_{gt}, \kappa_{gt})$ where $\mathscr{B}(\mu_{gt}, \kappa_{gt})$ is the beta distribution with mean $\mu_{gt} \in (0, 1)$ and precision parameter $\kappa_{gt} > 0$. We use the parameterization advocated by Ospina and Ferrari (2012) so that the density function of b_{gt} is,

$$f(b_{gt} \mid \mu_{gt}, \kappa_{gt}) = \frac{\Gamma(\kappa_{gt})}{\Gamma(\mu_{gt}\kappa_{gt})\Gamma((1 - \mu_{gt})\kappa_{gt})} b_{gt}^{\mu_{gt}\kappa_{gt}-1}(1 - b_{gt})^{(1-\mu_{gt})\kappa_{gt}-1}$$

with $\mathbb{E}(b_{gt}) = \mu_{gt}$ and $\mathbb{V}\mathrm{ar}(b_{gt}) = \mu_{gt}(1 - \mu_{gt})/(\kappa_{gt} + 1)$. We model μ_{gt} and κ_{gt} in the same way as δ_{gt} with,

$$\operatorname{logit}(\mu_{gt}) = \alpha_\mu + \boldsymbol{x}'_{gt}\boldsymbol{\beta}_\mu + \eta_{\mu g} + \phi_\mu \sum_{g' \in \mathscr{N}_g} d^{\phi_\mu}_{g'(t-1)} + \theta_\mu d^{\theta_\mu}_{g(t-1)} \tag{2.3}$$

$$\log(\kappa_{gt}) = \alpha_\kappa + \boldsymbol{x}'_{gt}\boldsymbol{\beta}_\kappa + \eta_{\kappa g} + \phi_\kappa \sum_{g' \in \mathscr{N}_g} d^{\phi_\kappa}_{g'(t-1)} + \theta_\kappa d^{\theta_\kappa}_{g(t-1)} \tag{2.4}$$

where, as in (2.2), the $\boldsymbol{\beta}$ parameters represent main effects for the covariates $\boldsymbol{x}_{gt}$, the η parameters represent spatially correlated random effects designed to capture the effect of unmeasured covariates, the ϕ parameters capture possible temporal effects of neighboring grid cells and the θ parameters capture the temporal dynamics of the grid cell itself.

2.2. Priors

We use vaguely informative priors for the ϕ and θ parameters. Specifically, we assume that all of the ϕ and θ parameters are *a priori* independent $\mathcal{N}(0, 10^2)$ random variates. Vague $\mathcal{N}(0, 10^2)$ priors are used for each of the α parameters.

For the $\boldsymbol{\beta}$ parameters, we desire to perform variable selection by learning the covariates that are important in explaining MPB damage and shrinking the remaining coefficients. We do not know *a priori* what variables to include in our model; therefore, we use the Bayesian LASSO prior of Park and Casella (2008) for $\boldsymbol{\beta}_\delta$, $\boldsymbol{\beta}_\mu$, and $\boldsymbol{\beta}_\kappa$, which constrains the coefficients to be shrunk toward zero. Specifically, for $\boldsymbol{\beta}_\delta$ we assume,

$$\begin{aligned}
\boldsymbol{\beta}_\delta &\sim \mathcal{N}_p\left(\boldsymbol{0}, \boldsymbol{D}_{\delta\tau}\right), \\
\boldsymbol{D}_{\delta\tau} &= \operatorname{diag}\left(\tau_{\delta 1}^2, \ldots, \tau_{\delta p}^2\right), \\
\tau_{\delta 1}^2, \ldots, \tau_{\delta p}^2 &\sim \prod_{i=1}^{p} \frac{\lambda_\delta^2}{2} \exp\left\{\frac{-\lambda_\delta^2 \tau_{\delta i}^2}{2}\right\} d\tau_{\delta i}^2, \\
f\left(\lambda_\delta^2\right) &\propto \left(\lambda_\delta^2\right) \exp\left\{-2\lambda_\delta^2\right\},
\end{aligned}$$

where $f(\cdot)$ denotes a density function. Equivalent priors were used for $\boldsymbol{\beta}_\mu$ and $\boldsymbol{\beta}_\kappa$.

In total, there are $3 \times 42 \times 55 = 6930$ η parameters which is a computational challenge. To help alleviate this problem, we use the sparse reparameterization of a conditional autoregressive model with dimension reduction as developed by Hughes and Haran (2013). Specifically, let $\boldsymbol{A}$ represent the 2310×2310 adjacency matrix of the grid cells with entries given by $\operatorname{diag}(\boldsymbol{A}) = \boldsymbol{0}$ and $\boldsymbol{A}_{ij} = 1$ if i and j are neighbors (share an edge) and 0 otherwise.

Let $\boldsymbol{M}$ be a $2310 \times q$ matrix of the first q columns of the Moran basis $\boldsymbol{P}^\perp \boldsymbol{A} \boldsymbol{P}^\perp$ where $\boldsymbol{P}^\perp = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, the projection onto the orthogonal column space of $\boldsymbol{X}$ and $\boldsymbol{X}$ is the $2310 \times P$ matrix of time-constant covariates. We set $\eta_g = \boldsymbol{m}_{(g)}' \boldsymbol{\eta}^\star$ where $\boldsymbol{m}_{(g)}$ is the g^{th} row of $\boldsymbol{M}$ and $\boldsymbol{\eta}^\star = (\eta_1^\star, \ldots, \eta_q^\star)'$ is a vector of coefficients. We use the prior described in Hughes and Haran (2013) which is derived from the intrinsic conditional autoregressive (ICAR) model for $\boldsymbol{\eta}_\delta^\star$, $\boldsymbol{\eta}_\mu^\star$, and $\boldsymbol{\eta}_\kappa^\star$. Specifically, the prior for $\boldsymbol{\eta}_\delta^\star$ is

$$p(\boldsymbol{\eta_\delta}^\star|\tau) \propto \tau^{q/2} \exp\left(-\frac{\tau}{2} \boldsymbol{\eta_\delta}^{\star'} \boldsymbol{Q}_s \boldsymbol{\eta_\delta}^\star\right)$$

where τ is a smoothing parameter and $\boldsymbol{Q}_s = \boldsymbol{M}'\boldsymbol{Q}\boldsymbol{M}$ where $\boldsymbol{Q} = \operatorname{diag}(\boldsymbol{A}\boldsymbol{1}) - \boldsymbol{A}$ and $\boldsymbol{1}$ is a vector of 1s. As shown in Hughes and Haran (2013), this sparse parameterization for $\{\eta_g\}$ has the effect of (i) alleviating confounding between the main effects (the $\boldsymbol{\beta}$'s) and spatial random effects by constraining spatial smoothing to the orthogonal column space of $\boldsymbol{X}$ and (ii) reducing the dimension of $(\eta_1, \ldots, \eta_{2310})'$ from 2310 to q. Based upon (Hughes and

Haran 2013) and from preliminary model fitting, we chose to use ≈10 % of the total spatial random effects, $q = 250$, as there was very little change in the estimates of the η parameters using $q > 250$. Here, only the time-constant covariates were used to construct the Moran basis, $\boldsymbol{M}$. We explored using a time-varying basis using the time-varying covariates but found that the basis functions changed very little over time. We admit, however, that while including the time-varying covariates did not seem to impact this analysis, we note that this may not be the case for all applications.

2.3. Statistical Inference and Prediction

Let $\boldsymbol{\theta}_z = (\alpha_\delta, \boldsymbol{\beta}_\delta, \boldsymbol{\eta}^\star_\delta, \phi_\delta, \theta_\delta, \{\tau^2_{\delta i}\}_i, \lambda_\delta)'$ and $\boldsymbol{\theta}_b = (\alpha_\mu, \alpha_\kappa, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\kappa, \boldsymbol{\eta}^\star_\mu, \boldsymbol{\eta}^\star_\kappa, \phi_\mu, \phi_\kappa, \theta_\mu, \theta_\kappa, \{\tau^2_{\mu i}\}_i, \{\tau^2_{\kappa i}\}_i, \lambda_\delta, \lambda_\kappa)'$ denote the vector of model parameters associated with $\{z_{gt}\}$ (the zero-augmented piece) and $\{b_{gt}\}$ (the beta piece). Furthermore, let $\mathscr{Z}_0 = \{(g, t) : z_{gt} = 0\}$ denote the set of indices where $y_{gt} \neq 0$. Given the stick-breaking weights $\{y_{gt} = (1 - z_{gt})b_{gt}\}_{gt}$, we have the following log-likelihood functions,

$$\mathscr{L}(\boldsymbol{\theta}_z) = \sum_{g,t} \left[z_{gt} \log(\delta_{gt}) + (1 - z_{gt}) \log(1 - \delta_{gt}) \right], \tag{2.5}$$

$$\mathscr{L}(\boldsymbol{\theta}_b) = \sum_{(g,t)\in\mathscr{Z}_0} \log \left[f(b_{gt} \mid \mu_{gt}, \kappa_{gt}) \right], \tag{2.6}$$

where the forms for δ_{gt}, μ_{gt}, and κ_{gt} are given in Eqs. (2.2), (2.3), and (2.4), respectively. The joint log-likelihood for $(\boldsymbol{\theta}_z, \boldsymbol{\theta}_b)$ is specified similarly to Ospina and Ferrari (2012) and is given by $\mathscr{L}(\boldsymbol{\theta}_z, \boldsymbol{\theta}_b) = \mathscr{L}(\boldsymbol{\theta}_z) + \mathscr{L}(\boldsymbol{\theta}_b)$. Due to the simple forms for $\mathscr{L}(\boldsymbol{\theta}_z)$ and $\mathscr{L}(\boldsymbol{\theta}_b)$ above, we opt to estimate $\boldsymbol{\theta}_z$ and $\boldsymbol{\theta}_b$ by drawing from their respective posterior distributions using a Gibbs sampler where we first draw $\boldsymbol{\theta}_z \sim f(\boldsymbol{\theta_b} \mid \boldsymbol{\theta}_z, \{y_{gt}\})$ then $\boldsymbol{\theta}_b \sim f(\boldsymbol{\theta}_b \mid \boldsymbol{\theta}_z, \{y_{gt}\})$.

The complete conditional distributions $f(\boldsymbol{\theta_b} \mid \boldsymbol{\theta}_z, \{y_{gt}\})$ and $f(\boldsymbol{\theta}_b \mid \boldsymbol{\theta}_z, \{y_{gt}\})$ are not available in closed form. Because of this, we use an adaptive Metropolis algorithm based on Haario et al. (2001) to update $\boldsymbol{\theta}_z$ and $\boldsymbol{\theta}_b$. Specifically, we use Gaussian proposal distributions where the variance of the proposal is set to be the variance of all previous draws. To obtain estimates of the parameters, we ran a chain for 1,000,000 iterations to ensure that the MCMC standard errors were small enough (Flegal et al. 2008).

An important component in this study is predicting what regions will be damaged (the z_{gt} component) and the amount of damage (the b_{gt} component) for the year $t^\star = 11$. To make predictions, we obtain draws of $y_{1t^\star}, \ldots, y_{Gt^\star}$ from the joint posterior predictive distribution using the identity,

$$\begin{aligned} \pi(z_{gt^\star}, b_{gt^\star}, \boldsymbol{\theta}_z, \boldsymbol{\theta}_b \mid \{y_{gt}\}_{gt}) = {} & \pi_{Z_{gt^\star}}(z_{gt^\star} \mid \boldsymbol{\theta}_z, \boldsymbol{\theta}_b, \{y_{gt}\}_{gt}) \pi_{B_{gt^\star}}(b_{gt^\star} \mid \boldsymbol{\theta}_z, \boldsymbol{\theta}_b, \{y_{gt}\}_{gt}) \\ & \times \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_z, \boldsymbol{\theta}_b \mid \{y_{gt}\}_{gt}), \end{aligned} \tag{2.7}$$

where we use π to denote a posterior distribution. From (2.7), we can obtain draws from the posterior predictive distribution of $\{(z_{gt^\star}, b_{gt^\star})\}$ by drawing $z_{gt^\star} \sim \text{Bern}(\delta_{gt^\star})$ and $b_{gt^\star} \sim \mathscr{B}(\mu_{gt^\star}, \kappa_{gt^\star})$ for each draw of $(\boldsymbol{\theta}_z, \boldsymbol{\theta}_b)$ obtained from the posterior distribution.

Draws from the posterior predictive distribution also give a measurement of the uncertainty associated with the prediction.

3. RESULTS

We consider $P = 5$ covariates to include as the vector $\boldsymbol{x}_{gt}$. For each grid cell g, we calculate the (i) August mean maximum temperature in degrees Celsius, (ii) January mean minimum temperature in degrees Celsius, (iii) mean annual precipitation in inches, (iv) terrain slope in percent rise, and (v) elevation in feet. Each of these components has been shown to have an impact on mountain pine beetle outbreaks in the western United States (see, for example,Waring and Pitman 1985; Mitchell and Preisler 1991; Negron and Popp 2004). Weather variables were taken from the PRISM dataset which is publicly available at. http://www.prism.oregonstate.edu/ The PRISM data estimate monthly weather data over a contiguous grid at a resolution of 0.0416 decimal degrees latitude and longitude ($\sim$4 km) cells (Daly et al. 2002) and align with the resolution of our gridded MPB data. The weather variables were adjusted to account for a one-year lag between infestation and the time MPB damage is detected in the ADS. That is, for August mean maximum and mean annual precipitation we used data from 1999 to 2008 and for January mean minimum temperatures we used data from 2000 to 2009. Slope and elevation data for each site were generated from a Digital Elevation Map (DEM) of the state of Colorado in ArcGIS, where slope is calculated based upon the maximum rate of change in elevation over the distance from one site and its neighboring sites.

3.1. Model Selection

Intuitively, for a grid cell g, the ϕ parameters are associated with the covariate $\sum_{g' \in \mathcal{N}_g} d^{\phi}_{g'(t-1)}$ and represent an added effect due to the cumulative damage to neighbors of g at the previous time period. Similarly, the θ parameters are associated with the covariate $d^{\theta}_{g(t-1)}$ and represent an added effect due to damage at grid cell g but at the previous time period. The model postulated in (2.2), (2.3), and (2.4) requires a choice for the covariates $d^{\phi}_{gt} \in \{y_{gt}, \widetilde{y}_{gt}\}$ and $d^{\theta}_{gt} \in \{y_{gt}, \widetilde{y}_{gt}\}$. The question, then, is which measure of damage (y or $\widetilde{y}$) is a better predictor of MPB damage? To answer this question, we fit the proposed model for each combination of $d^{\phi_\delta}_{gt}$, $d^{\theta_\delta}_{gt}$, $d^{\phi_\mu}_{gt}$, $d^{\theta_\mu}_{gt}$, $d^{\phi_\kappa}_{gt}$, and $d^{\theta_\kappa}_{gt}$ (totaling $2^6 = 64$ models).

In this particular application, predictive performance was the most important because the ultimate goal is to predict the likelihood that MPB's will appear in a particular grid cell allowing subsequent intervention strategies to be made. To assess prediction accuracy, we left out the year $t = 2010$ and compared model predictions of the occurrence of damage ($z_{gt^\star}$) and amount of damage ($b_{gt^\star}$) to the observed $\widetilde{y}_{gt}$. We compared each model's prediction of the hold-out sample based on the misclassification rate, root mean square prediction error (RMSPE), and continuous ranked probability score (CRPS; Gneiting and Raftery 2007). The misclassification rate was defined as $(2310^{-1}) \sum_{g=1}^{2310} \mathbb{1}_{\{\ddot{z}_{gt^\star} \neq z_{gt^\star}\}}$ where $\ddot{z}_{gt^\star} = 1$ if $\mathbb{P}\mathrm{r}(z_{gt^\star} = 1 \mid \{y_{gt}\}_{gt}) \geq 0.5$ and $\ddot{z}_{gt^\star} = 0$ otherwise. The RMSPE is calculated as

Table 1. Top 5 models, ranked according to percent misclassification low to high and DIC for various choices of d in (2.2), (2.3), and (2.4).

$d_{gt}^{\phi_\delta}$	$d_{gt}^{\theta_\delta}$	$d_{gt}^{\phi_\mu}$	$d_{gt}^{\theta_\mu}$	$d_{gt}^{\phi_\kappa}$	$d_{gt}^{\theta_\kappa}$	DIC	Misclass	RMSPE	CRPS
y_{gt}	y_{gt}	$\widetilde{y}_{gt}$	y_{gt}	$\widetilde{y}_{gt}$	y_{gt}	0.00	0.234	0.085	51.591
y_{gt}	y_{gt}	y_{gt}	$\widetilde{y}_{gt}$	y_{gt}	y_{gt}	479.741	0.234	0.088	52.626
y_{gt}	y_{gt}	y_{gt}	$\widetilde{y}_{gt}$	$\widetilde{y}_{gt}$	y_{gt}	457.148	0.237	0.083	50.652
y_{gt}	y_{gt}	y_{gt}	$\widetilde{y}_{gt}$	y_{gt}	$\widetilde{y}_{gt}$	432.706	0.240	0.085	52.053
y_{gt}	y_{gt}	y_{gt}	$\widetilde{y}_{gt}$	$\widetilde{y}_{gt}$	$\widetilde{y}_{gt}$	448.334	0.242	0.083	51.028

$\sqrt{(2310^{-1})\sum_{g=1}^{2310}(\widehat{y}_{gt^\star} - y_{gt})^2}$ where $\widehat{y}_{gt}$ is the posterior predictive mean of y_{gt}. Because the CRPS is only defined for continuous variables, we calculate CRPS only for those $b_{gt^\star}$ for which $z_{gt^\star} = 0$. That is, we calculate the CRPS of the random variable $b_{gt^\star} \mid z_{gt^\star} = 0$. As the RMSPE and CRPS differences were relatively small we chose the model based upon percent misclassification. As an additional measure, although secondary to predictive performance, we compared each model's fit based on the deviance information criterion (DIC) of Spiegelhalter et al. (2002).

Table 1 displays the top 5 models ranked in terms of misclassification rate. From Table 1, we note that the model which had the lowest misclassification rate also has the lowest DIC (we adjusted the DIC values so that the minimum observed DIC value was 0). For example, a DIC value of 479.741 means that the DIC value was 479.741 greater than the first model in Table 1. If we rank models according to DIC, the results in Table 1 change. Other than the first model in Table 1, the next best models according to DIC had misclassification rates greater than 27 %.

Considering the best model in Table 1, the $\widetilde{y}_{gt}$ are preferred for the ϕ_μ and ϕ_κ coefficients but not the ϕ_δ coefficient. This result seems to suggest that the cumulative amount of damage ($\widetilde{y}_{gt}$) to neighbors of grid cell g is predictive of the amount of damage but not of the occurrence of damage. Rather, the occurrence of damage is better explained and predicted by the amount of damage (y_{gt}) incurred at neighboring grid cells in the previous time period.

We note that, in model (2.2), $d_{gt}^{\phi_\mu}$ and $d_{gt}^{\theta_\mu}$ were found to be different covariates. As it is not fully known how MPBs migrate between spatial locations we allowed the covariates to differ for the $d_{gt}^{\theta_\mu}$ and $d_{gt}^{\phi_\mu}$ components in the model. Doing so allowed us to explore the differences between within-cell and between-cell temporal correlations in the data. That is, cumulative damage done to neighboring grid cells seems to be more explanatory of the amount of damage in a grid cell than cumulative damage done within a grid cell at the previous time period.

Prior to concluding this section, we note that comparing the different models based on RMSPE and CRPS was more challenging because the observed spread of RMSPE and CRPS between models was small. For example, the best model according to RMSPE had RMSPE $= 0.081$ compared to a maximum RMSPE of 0.095. This suggests that our model is able to predict the z_{gt} component better than the b_{gt} component.

3.2. Model Fit Results

For the best model (row 1 in Table 1), Table 2 displays posterior summaries (medians and 95 % credible intervals) of the main effect coefficients $\boldsymbol{\beta}$, ϕ, and θ in each of (2.2), (2.3), and (2.4). Values represent the percent change in the odds ratio (for δ and μ) or the percent change in κ. For rows 1 through 4, values indicate percent change due to a unit increase in the covariate. For row 5, values indicate percent change given a 1000 foot increase in elevation. For rows 6 and 7, values indicate percent change given 10 % increase in damage.

As expected, several of the parameters in (2.2) and (2.3) have opposite signs for the same covariate. For example, when the amount of precipitation increases (i) the mean amount of MPB damage, given damage occurred, decreases and (ii) the probability of no damage increases. This result is also true for August mean maximum temperatures. That is, the data indicate that when August maximum temperatures increase (i) the mean amount of MPB damage decreases and (ii) the probability of no damage increases. The data also show the higher the August temperature the more variable the amount of damage is.

The result that the mean amount of MPB damage decreases with increases in August temperature is opposite from previous studies. That is, previous studies by Negron and Popp (2004); Zhu et al. (2008) show increases in temperature lead to greater damage. As multicollinearity may be a matter of concern, we tested for multicollinearity by assessing the correlation among the variables magnitude less than 0.65. We also removed each of the covariates out of the model one at a time. However, the signs for the covariates stayed the same across models suggesting this result is not due to multicollinearity. We hypothesize that this contradiction of previous results occurred because we used monthly rather than daily average temperatures. We hypothesize that this contradiction of previous results occurred because we used monthly rather than daily average temperatures. That is, because MPB populations are diminished with multiple days of extreme cold temperatures (e.g., less than −30 °C), using monthly temperatures it causes the extreme cold or heat events days to be masked. However, further exploration into this result is needed.

In terms of the landscape effects, elevation has a positive relationship with the mean amount of MPB damage and a negative relationship with the probability of no damage as expected. However, the effect of slope on δ and μ has the same sign. That is, slope has a positive relation to mean MPB damage (μ) and the probability of no damage (δ). This result also seems opposite of what intuition might imply. For example, we *a priori* might expect that as the slope increases, δ increases whereas μ decreases. Upon closer inspection, this opposing relation can be explained by the diversity of the tree stands. Because only certain types of trees are able to grow on steep slopes, the probability of no damage increases with slope because MPBs might not infest these type of trees. However, the amount of damage at these high slopes can be more substantial because, conditional on damage occurring, the type of tree within the grid cell is not resilient against MPB damage.

For ϕ, note that as the cumulative damage to the neighbors of a grid cell in the previous year decreases the probability of MPB damage within the grid cell increases (δ decreases). Additionally, as the neighbors of a grid cell become less damaged, the damage within the grid cell can be more substantial. This result seems to suggest that MPBs tend to consume the resources in neighboring cells before consuming the resources within that grid cell.

Table 2. Posterior medians and credible intervals for main effect coefficients $\boldsymbol{\beta}$, ϕ, and θ in each of (2.2), (2.3), and (2.4).

	δ		μ		κ	
Covariate	Median	95 % CI	Median	95 % CI	Median	95 % CI
Jan. Temp	0.05	(−0.06, 1.44)	−0.83	(−1.56, −0.06)	−1.50	(−2.54, −0.51)
Aug. Temp	15.34	(13.41, 16.54)	−8.74	(−9.64, −7.79)	9.40	(8.36, 10.46)
Precip	1.46	(0.87, 1.88)	−0.17	(−0.46, 0.07)	1.01	(0.70, 1.36)
Slope	2.70	(1.30, 4.03)	8.35	(7.55, 9.14)	−13.37	(−14.45, −12.47)
Elev	−13.24	(−17.62, −11.97)	5.19	(0.63, 9.25)	3.64	(−0.77, 8.08)
ϕ	−7.35	(−7.78, −6.87)	−0.18	(−0.21, −0.13)	0.15	(0.11, 0.18)
θ	−14.04	(−14.61, −13.51)	3.02	(2.83, 3.25)	−0.81	(−0.87, −0.76)

Values represent the percent change in the odds ratio (for δ and μ) or the percent change in κ. For rows 1 through 4, values indicate percent change due to a unit increase in the covariate. For row 5, values indicate percent change given a 1000 foot increase in elevation. For rows 6 and 7, values indicate percent change given 10 % increase in damage.

Hence, if a grid cell's neighbors have been nearly entirely damaged, MPBs will finish off the resources within a grid cell before migrating to regions with more undamaged trees.

Finally, for θ, the temporal effect of the MPB damage in the previous year, as the percent of MPB damage from the previous year increases the probability of no MPB damage (δ) decreases and the mean amount of damage (μ) increases. This result is expected and supports the idea that MPBs consume the resources within the grid cell before migrating to other regions with undamaged trees. In the ADS survey data, if an area is designated as highly damaged subsequent years of data collection focus on regions near areas that have been previously impacted, capturing the nature of how MPBs migrate.

3.3. Predictive Results

Figure 3a displays a map of the estimated probability $(1 - \delta_{gt^\star})$ for $t^\star = 11$ (the year 2011). That is, Fig. 3a displays the posterior predictive probability of MPB damage across the study region. Historically, from the trend seen in Fig. 2, the MPB damage is moving predominantly in an easterly direction. While our model continues to predict more damage to the east, noticeably, our model is predicting damage in 2011 to be more in a south-eastern direction. That is, from Fig. 3 we see a high probability of MPB damage in the south-eastern region of the study area.

Figure 3b displays the posterior median of the posterior predictive distribution of $y_{gt^\star}$ given $z_{gt^\star} = 0$ (the amount of damage conditional on MPB damage occurring in the area). That is, Fig. 3b displays the median amount of MPB damage predicted given that MPB damage occurred. Figure 3b shows that we predict the highest amount of damage to be in the south-eastern region. Combining Fig. 3a, b, we are able to highlight regions wherein intervention strategies are, perhaps, most effective.

Figure 4 displays the full posterior predictive distribution of $y_{gt^\star}$ for three randomly selected grid cells. Based upon inspection of a larger sample, we found that the majority of predictive distributions follow this pattern—namely, a peak near zero with a heavy tail

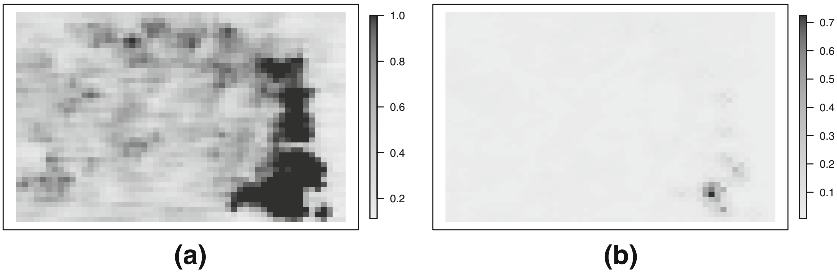

Figure 3. **a** Predicted probability of MPB damage ($1-\delta_{gt^{\star}}$) for $t^{\star}=11$ and **b** posterior median of the predicted damage $y_{gt^{\star}}$ given $z_{gt^{\star}}=0$.

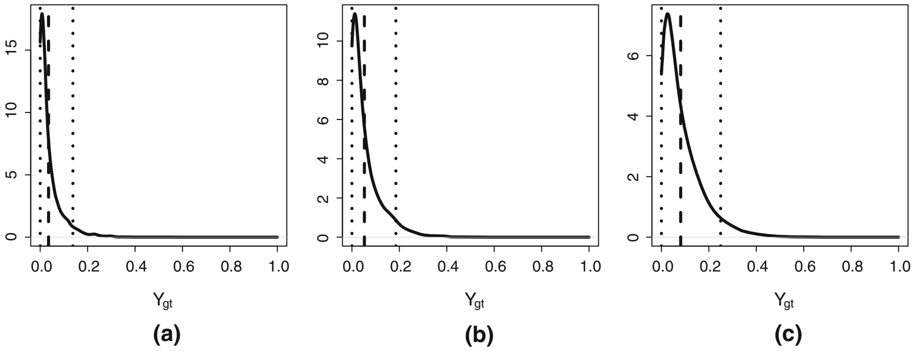

Figure 4. Posterior predictive distributions of $y_{gt^{\star}}$ for three randomly selected grid cells. *Vertical dashed lines* denote the posterior predictive mean and the *dotted lines* denote a 95 % highest posterior density interval.

(although the heaviness of the tail changes from grid cell to grid cell). This pattern suggests that large pine beetle damage is possible but not likely.

4. DISCUSSION AND CONCLUSIONS

This article focused on modeling and predicting aerial detection survey data of MPB damage in Colorado. Specifically, we used a stick-breaking representation to enforce monotonicity constraints of the cumulative damage to grid cells in a region in Colorado. We model the resulting stick-breaking weights using a zero-inflated beta regression model wherein the probability of zero damage as well as the mean and dispersion of a beta distribution vary over space and time.

The ultimate goal of this work was to build a predictive model for MPB damage using ADS data. We demonstrated this predictive methodology in Sect. 3.3 by the ability to predict regions where the highest amount of damage will occur. We note that the predictive distributions in Fig. 4 show a very heavy tail. This heavy-tailed nature of MPB damage challenges our assumption of a beta distribution for positive damage. An alternative modeling strategy could include the use of extreme value distributions to more appropriately model the upper tail of MPB damage.

For the ADS data, there are portions of the region wherein no trees are present; hence, there can be no MPB damage to these grid cells. However, the ADS data do not directly distinguish between zero damage resulting from no trees in a grid cell and a “true” MPB damage of zero. Due to the large amount of MPB damage in the spatial region of interest, it may be safe to assume that if a grid cell is never damaged then no trees are present. However, this judgement may be uncertain so we do not wish to simply throw out grid cells for which there is never any damage. Future work on this project should include development of methodology to appropriately partition the spatial region into regions for which MPB damage is possible.

As a final note, one area that requires more attention is the temporal lag structure used in the model. That is, in this work we accounted for the one-year lag structure between impact and detection of the infestation. However, the use of other lag structures, such as distributed lag models, may give more understanding to the temporal dynamics of MPB damage.

ACKNOWLEDGEMENTS

The authors wish to thank the Colorado State Forest Service for providing the data and describing the process of aerial surveying. The majority of this work was performed while K. A. Kaufeld was a graduate student at University of Northern Colorado and a student visitor at the National Center for Atmospheric Research. M. J. Heaton was a postgraduate scientist at the National Center for Atmospheric Research. The National Center for Atmospheric Research is managed by the University Corporation for Atmospheric Research under the sponsorship of the National Science Foundation.

[ReceivedOctober 2013. AcceptedSeptember 2014. Published OnlineDecember 2014.]

REFERENCES

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, 99–113.

Ferrari, S. L. P. and Cribari-Neto, F. (2004), “Beta regression for modelling rates and proportions”, *Journal of Applied Statistics*, 31, 799–815.

Flegal, J. M., Haran, M., and Jones, G. L. (2008), “Markov chain monte carlo: Can we trust the third significant figure?” *Statistical Science*, 23, 250–260.

Gneiting, T. and Raftery, A. E. (2007), “Strictly proper scoring rules, prediction, and estimation”, *Journal of the American Statistical Association*, 102, 359–378.

Haario, H., Saksman, E., and Tamminen, J. (2001), “An adaptive metropolis algorithm”, *Bernoulli*, 7, 223–242.

Harris, J. L. (2003), “Forest insect and disease conditions in the rocky mountain region, 2002”, *USDA For. Serv., Rocky Mountain Region, Renewable Resources*, 28 pp.

Harris, J. L., Mask, R., and Witcosky, J. (2002), “Forest insect and disease conditions in the rocky mountain region, 2000–2001”. *USDA For. Serv., Rocky Mountain Region, Renewable Resources*, 42 pp.

Hatfield, L. A., Boye, M. E., Hackshaw, M. D., and Carlin, B. P. (2012), “Multilevel bayesian models for survival times and longitudinal patient-reported outcomes with many zeros”, *Journal of the American Statistical Association*, 107, 875–885.

Hughes, J. and Haran, M. (2013), “Dimension reduction and alleviation of confounding for spatial generalized linear mixed models”, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 139–159.

Kieschnick, R. and McCullough, B. D. (2003), “Regression analysis of variates observed on (0,1): percentages proportions and fractions”, *Statistical Modelling*, 3, 193–213.

Mitchell, R. G. and Preisler, H. K. (1991), “Analysis of spatial patterns of lodgepole pine attacked by outbreak populations of the mountain pine beetle”, *Forest Science*, 37, 1390–1408.

Negron, J. F. and Popp, J. B. (2004), “Probability of ponderosa pine infestation by mountain pine beetle in the colorado front range”, *Forest Ecology and Management*, 191, 17–27.

Ospina, R. and Ferrari, S. L. (2010), “Inflated beta distributions”, *Statistical Papers*, 51, 111–126.

——— (2012), “A general class of zero-or-one inflated beta regression models”, *Computational Statistics and Data Analysis*, 56, 1609–1623.

Park, T. and Casella, G. (2008), “The bayesian lasso”, *Journal of the American Statistical Association*, 103, 681–686.

Raffa, K. F., Aukema, B. H., Carroll, A. L., Hicke, M. G., Turner, M. G., and Romme, W. H. (2008), “Cross-scale drivers of natural disturbances prone to anthropogenic amplification: the dynamics of bark beetle eruptions”, *BioScience*, 58, 501–517.

Sethuraman, J. (1994), “A constructive definition of dirichlet priors”, *Statistica Sinica*, 4, 639–650.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), “Bayesian measures of model complexity and fit”, *Journal of the Royal Statistical Society: Series B*, 64, 583–639.

Waring, R. H. and Pitman, G. B. (1985), “Modifying lodgepole pine stands to change susceptibility to mountain pine beetle attack”, *Ecology*, 66, 889–897.

Wieczorek, J. and Hawala, S. (2011), “A Bayesian zero-one inflated beta model for estimating poverty in us counties”, *JSM Proceedings*, Section on Survey Research Methods, Miami, FL, 2812–2822.

Wieczorek, J., Nugent, C., and Hawala, S. (2012), “A Bayesian zero-one inflated beta model for small area shrinkage estimation”, *JSM Proceedings*, Section on Survey Research Methods, San Diego, CA, 3896–3910.

Williams, D. W. and Liebhold, A. M. (2002), “Climate change and the outbreak ranges of two north american bark beetles”, *Agricultural and Forest Entomology*, 4, 87–99.

Zheng, Y. and Zhu, J. (2008), “Markov chain monte carlo for a spatial-temporal autologistic regression model”, *Journal of Computational and Graphical Statistics*, 17, 123–137.

Zhu, J., Huang, H. C., and Wu, J. (2005), “Modeling spatial-temporal binary data using markov random fields”, *Journal of Agricultural, Biological, and Environmental Statistics*, 10, 212–225.

Zhu, J., Zheng, Y., Carroll, A. L., and Aukema, B. H. (2008), “Autologistic regression analysis of spatio-temporal binary data via monte carlo maximum likelihood”, *Journal of Agricultural, Biological, and Environmental Statistics*, 13, 84–98.