Environ Ecol Stat (2006) 13:409–426
DOI 10.1007/s10651-006-0020-x

ORIGINAL ARTICLE

A spatial zero-inflated poisson regression model for oak regeneration

Stephen L. Rathbun · Songlin Fei

Received: May 2004 / Revised: December 2004

Abstract Ecological counts data are often characterized by an excess of zeros and spatial dependence. Excess zeros can occur in regions outside the range of the distribution of a given species. A zero-inflated Poisson regression model is developed, under which the species range is determined by a spatial probit model, including physical variables as covariates. Within that range, species counts are independently drawn from a Poisson distribution whose mean depends on biotic variables. Bayesian inference for this model is illustrated using data on oak seedling counts.

Keywords Bayesian hierarchical spatial Model · MCMC algorithm · Spatial probit model

1 Introduction

Ecological surveys often involve counts of the numbers of individuals of one or more species at sample sites scattered throughout a study region. Their intent may be to obtain a better understanding of what environmental factors or habitat conditions are favorable to the species of interest. If the locations of individual organisms are realized from a spatial inhomogeneous Poisson process (Cressie 2001) whose log intensity is a linear function of known environmental variables, then by the independent increments property of the Poisson process, the counts at different locations

S. L. Rathbun(✉)
Department of Health Administration, Biostatistics and Epidemiology, University of Georgia,
Athens, GA 30605, USA
e-mail: rathbun@uga.edu

S. Fei
Department of Forestry,
University of Kentucky
Lexington, KY 40546, USA
e-mail: songlin.fei@uky.edu

are independently distributed. Poisson regression may be then used to model the effects of environmental variables on species abundances. In practice, however, not all pertinent variables are included in the analysis. This may result not only in overdispersion, but also spatial dependence in the counts data. Approaches to analyzing spatially dependent counts data include the introduction of random effects (Diggle et al. 1998), or marginal modeling using generalized estimating equations to estimate model parameters (Gotway and Stroup 1997; Gotway and Wolfinger 2003).

Ecological counts data often include an excess of zeros (e.g., Welsh et al. 2000; Leathwick and Austin 2001; O'Neill and Faddy 2003) owing either to the inclusion of habitat unsuitable to the species, or to the limited ability of the species to disperse into all parts of the study region. Lamber (1992) introduced the zero-inflated Poisson regression model to account for an excess of zeros in counts of manufacturing defects. Zero-inflated Poisson regression has been applied to model the numbers of sightings of a rare possum species (Welsh et al. 1996), numbers of insect pests on sugarcane (Vieira et al. 2000), and numbers of whiteflies on experimentally manipulated poinsettias (Van Iersel et al. 2000, 2001). Zero-inflated negative binomial regression has been proposed to model overdispersed data (Welsh et al. 1996). Hall (2000) introduced zero-inflated Poisson models with random effects for applications in longitudinal data analysis.

Agarwal et al. (2002) proposes a zero-inflated model for spatial count data. In accordance with the approach of Lambert 1992, they generate a zero with probability p, and data from a Poisson distribution with probability $1 - p$. Logistic regression is used to model the probabilities of the excess zeros, while the log linear model is used for the Poisson mean. Spatial dependence is introduced by adding spatially dependent random effects to the logistic regression and/or log linear models. Conditional on those random effects, excess zeros are generated independently. Consequently, any region, no matter how small, will contain an infinite number of sites with excess zeros, a pattern that is not compatible with the notion that excess zeros arise from the inclusion of regions that are unsuitable to the species.

This paper introduces a zero-inflated Poisson model in which the excess zeros are generated by a spatial probit model (Heagerty and Lele 1998). Under this model, an excess zero is generated at a given site if the realization of a Gaussian random field falls below a threshold. The collection of sites exceeding the threshold form a random set, taken here to be habitat suitable to the species of interest. Here, the realization of the random field is interpreted to be a measure of habitat suitability. By letting the mean of the random field be a linear function of covariates, the effects of environmental variables on habitat suitability can be modeled. Within suitable habitat, counts are generated according to a Poisson distribution. The log Poisson mean is taken to be a linear function of environmental covariates.

The proposed zero-inflated Poisson regression model is developed in Sect. 3, and Bayesian inference for the model parameters is considered in Sect. 4. Sect. 5 illustrates inferences for the spatial zero-inflated Poisson regression model using data on oak regeneration. The oak regeneration data are described in Sect. 2.

2 Oak regeneration data

Throughout eastern North America, natural regeneration of oaks (*Quercus* species) is often difficult to obtain even where oaks are the dominant components of the overstory before harvest. Although this problem is widespread both geographically

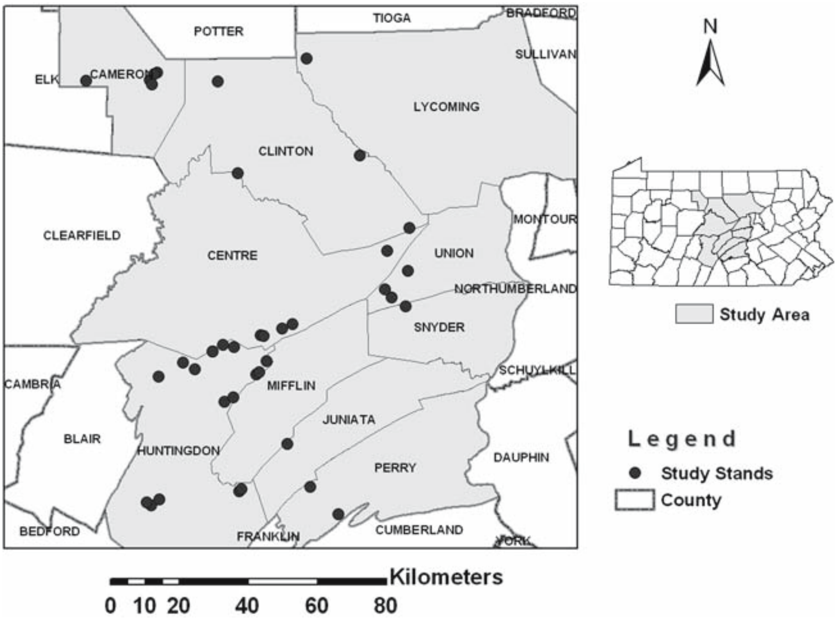

Fig. 1 Locations of the 38 mixed-oak stands in central Pennsylvania

and among species, there is no universal solution. The major causes of regeneration difficulty may change from site to site and from region to region. Because of this variability, the "oak regeneration problem" is perceived to have both local and regional aspects (Lorimer 1992). Specific prescriptions of local conditions necessary for oak regeneration are required to supplement general guidelines for regenerating oaks (Crow 1988). In an effort to understand and to help natural oak regeneration, a long term Oak Regeneration Project has been conducted in central Pennsylvania since 1996.

We shall explore the oak regeneration problem using data from 38 mixed-oak stands surveyed in central Pennsylvania from 1996 to 2000 (Fig. 1). All stands were surveyed 1 year prior to harvest. Within each stand, depending on stand size, 15–39 permanent circular plots, 8.02 m radius (20th-acre) each, were placed systematically in an approximately square grid. Fig. 2 shows the layout of the plots in a typical stand. Complete data were available for 1,331 plots. Oak regeneration was enumerated within four milacre (1.13 m radius) subplots established within each plot. This investigation focuses on three oak species, prevalent in central Pennsylvania: chestnut oak (*Quercus prinus*), white oak (*Q. alba*), and northern red oak (*Q. rubra*).

Both physical and biotic variables were used to explain variation in oak regeneration. Physical variables include elevation, slope shape (sum of percent slope uphill, downhill, and at 90° to aspect), slope percentage, slope aspect, and exposure angle (the angle between the visible east and west horizons). Biotic variables include canopy oak density by species, percent cover of hayscented ferns (*Dennstaedtia punctilobula*), and percent cover of heather shrubs (blueberry (*Vaccinium* species) and huckleberry (*Gaylussacia baccata*)).

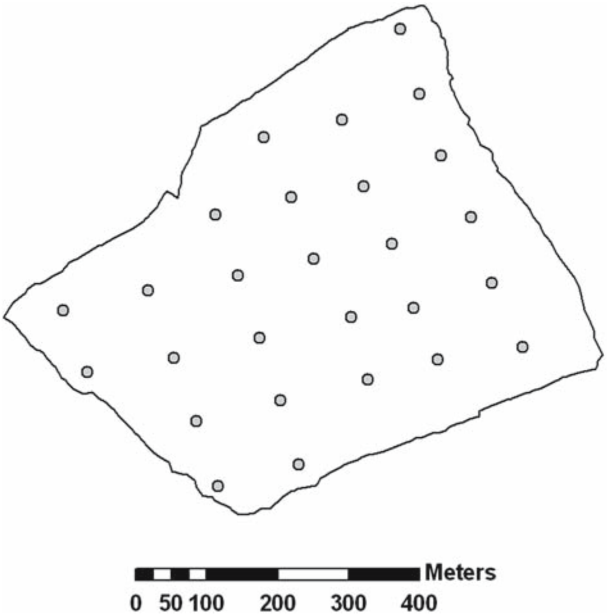

Fig. 2 Layout of plots in a typical stand

3 Zero-inflated Poisson model

The following defines a zero-inflated Poisson regression model in which the excess zeros are generated by a spatial probit model (Heagerty and Lele 1998). The multivariate probit model was introduced by Ashford and Sowden (1970), and is generated by first selecting a random vector from a multivariate Gaussian distribution. Then zeros are generated corresponding to elements falling below a threshold. A spatial version of the probit model is obtained by letting the elements of the variance–covariance matrix depend on the distance between the pair of sites.

Define the random field

$$Y(\mathbf{s}) = \alpha' \mathbf{x}_1(\mathbf{s}) + \varepsilon(\mathbf{s}),$$

where $\mathbf{x}_1(\mathbf{s})$ denotes a $p_1 \times 1$ vector of covariates observed at a location $\mathbf{s} \in D \subset \Re^2$, α is a $p_1 \times 1$ vector of parameters, and $\varepsilon(\cdot)$ is a zero-mean, Gaussian random field with covariance function $\tau^2 \rho(r;\gamma)\,; r \geq 0$ that depends on unknown parameters τ^2 and γ. Then the spatial probit model is defined by the binary random field

$$W(\mathbf{s}) = I(Y(\mathbf{s}) > \zeta)\,; \mathbf{s} \in D \subset \Re^2.$$

Here, the set $\{\mathbf{s} \in D : Y(\mathbf{s}) > 0\}$ is interpreted to be the habitat suitable to the species of interest. Not all of the parameters of this model are identifiable (De Oliveira 2000). Therefore, we shall fix the variance $\tau^2 = 1$ and threshold $\zeta = 0$.

To ensure that realizations of the Gaussian field $\varepsilon(\cdot)$ are continuous in mean square (see Stein (1999, pp. 20–22) for a formal definition), the correlation function $\rho(\cdot;\gamma)$

should be continuous at lag zero. The Matérn (1960) class of correlation functions

$$\rho\,(r;\gamma,\nu) = \frac{2\,(\gamma r/2)^{\nu}\,K_{\nu}\,(\gamma r)}{\Gamma\,(\nu)}, \tag{1}$$

where $K_{\nu}\,(\cdot)$ is the modified Bessel function of the second kind (Abramowitz and Stegun 1965), includes a parameter ν that controls the smoothness of the realizations of the random filed $\varepsilon\,(\cdot)$ (Stein 1999). Given that realizations of the random field $Y\,(\cdot)$ are not observable, however, the parameter ν cannot be identified from the data. Therefore, this parameter cannot be left free, but should be fixed at an appropriate value.

The spatial zero-inflated Poisson regression model is obtained by taking the count $Z\,(\mathbf{s}) = 0$ if $Y\,(\mathbf{s}) < 0$, and selecting $Z\,(\mathbf{s})$ from a Poisson distribution with mean

$$\lambda\,(\mathbf{s};\beta) = \exp\left\{\beta'\mathbf{x}_2\,(\mathbf{s})\right\}$$

if $Y\,(\mathbf{s}) > 0$. The $p_2 \times 1$ vector of covariates $\mathbf{x}_2\,(\mathbf{s})\,;\,\mathbf{s} \in D \subset \Re^2$ is used to model the effects of environmental variables on species abundance. The two collections of explanatory variables $\mathbf{x}_1\,(\cdot)$ and $\mathbf{x}_2\,(\cdot)$ may share covariates and may or may not be identical (Lambert 1992).

3.1 Inferential issues

Suppose the data

$$\{Z\,(\mathbf{s}_i), \mathbf{x}_1\,(\mathbf{s}_i), \mathbf{x}_2\,(\mathbf{s}_i) : i = 1,\ldots,n\}$$

are sampled at n sites in the study region D. To simplify notation, we shall take $z_i = Z\,(\mathbf{s}_i)\,, \mathbf{x}_{1i} = \mathbf{x}_1\,(\mathbf{s}_i)\,$, and $\mathbf{x}_{2i} = \mathbf{x}_2\,(\mathbf{s}_i)\,; i = 1,\cdots,n$. Then the joint distribution of $\mathbf{z}$ and $\mathbf{y}$ is given by

$$\begin{aligned} p\left(\mathbf{z},\mathbf{y}|\theta\right) &= (2\pi)^{-n/2}\left|\Sigma_{\gamma}\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}_1\alpha\right)'\Sigma_{\gamma}^{-1}\left(\mathbf{y}-\mathbf{X}_1\alpha\right)\right\} \\ &\quad\times\prod_{i=1}^{n}\left\{I\,(Y_i<0)\,I\,(z_i=0)+I\,(Y_i>0)\,\frac{1}{z_i!}\exp\left\{z_i\left(\beta'\mathbf{x}_i\right)-e^{\beta'\mathbf{x}_i}\right\}\right\}, \end{aligned} \tag{2}$$

where the $n \times p_1$ design matrix $\mathbf{X}_1$ has elements $x_{1j}\,(\mathbf{s}_i)\,$, the $n \times n$ correlation matrix Σ_{γ} has elements $\rho\left(\mathbf{s}_i - \mathbf{s}_j;\gamma\right)$. The likelihood is then obtained by integrating with respect to the unobserved random field **y:**

$$p\,(\mathbf{z}|\theta) = \int_{\Re^n} p\left(\mathbf{z},\mathbf{y}|\theta\right)\mathrm{d}\mathbf{y}. \tag{3}$$

If the latent random field $Y\,(\cdot)$ was observable, then maximum likelihood or Bayesian estimators of the parameters may be readily obtained from (2). Owing to the high dimension of the integral in (3), obtaining such estimators from the data alone is not straightforward. One approach is to reduce the dimension by adopting a composite likelihood approach. For example, Heagerty and Lele (1998) use pair-wise composite likelihood to estimate the parameters of their spatial probit model. Composite likelihood estimators are not as efficient (asymptotically) as maximum likelihood estimators. Moreover, the composite likelihood equations for our model cannot be simply expressed as a quadratic function of the data as in Heagerty and Lele, and the complex structure of the matrix of second derivatives makes it difficult to prove the consistency and asymptotic normality of the composite likelihood estimators as the area of the study region increases.

An alternative approach is to augment the data by simulating realizations of the random field $Y(\cdot)$, conditional on the data vector $\mathbf{z}$, and candidate values of the parameter vector θ. Chib and Greenberg (1998) employ this strategy to obtain Bayesian estimators and maximum likelihood estimators via the EM algorithm for the parameters of their multivariate probit model. This approach has also been used by Weir and Pettitt (1999, 2000) and De Oliveira (2000) for their spatial probit models. Since the large-sample inferential properties of the maximum likelihood estimator are yet to be proved, the following shall adopt a Bayesian approach to parameter estimation.

3.2 Prior choice

Little information is available for the elicitation of priors in the present application. Therefore, an objective Bayesian approach shall be taken here, but to ensure that the posterior distribution is proper, only proper priors shall be considered. Given the large sample size, it is expected that the data will dominate the prior. To simplify the implementation of the Monte Carlo Markov chain algorithm used to sample from the posterior distribution (Sect. 4), we shall consider the separable prior structure

$$\pi(\theta) = \pi(\alpha)\,\pi(\beta)\,\pi(\gamma).$$

Zellner's (1986) g-prior shall be adopted for α and β. Take $\alpha \sim N(\alpha_0, \mathbf{V}_\alpha)$ and $\beta \sim N(\beta_0, \mathbf{V}_\beta)$, where $\mathbf{V}_\alpha = g_\alpha \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}$ and $\mathbf{V}_\beta = g_\beta \left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}$. The prior variance–covariance matrices of these priors take the variability of each of the explanatory variables into account, yielding smaller prior variances for explanatory variables that show greater variability. Specifying large values of g_α and g_β yield noninformative priors. In the present application, we take $g_\alpha = g_\beta, = 1000$. To favor the null model in which the counts are independent of the explanatory variables, we shall take $\alpha_0 = \mathbf{0}$ and $\beta_0 = \mathbf{0}$. Zero intercept terms correspond to a model in which the range of the species covers half of the study region, and in which the mean counts are equal to one within that range.

The choice of prior for the parameter γ of the correlation function depends on the parametric form for that correlation function. Here, we shall take $\rho(r;\gamma) = \gamma^r$ and γ to be sampled from the Beta(a, b) distribution

$$\pi(\gamma) = \frac{\Gamma(a,b)}{\Gamma(a)\,\Gamma(b)}\gamma^{a-1}(1-\gamma)^{b-1}$$

Note that this correlation function corresponds to a reparameterization of the exponential correlation function, which is obtained by taking $\nu = 0.5$ in expression (1). Initially, we shall take $a = b = 1$, yielding a uniform prior for γ. Note that the uniform prior is not noninformative in this case, but favors a correlation function that is halved with every unit increase in the distance between the pair of sites. In the present application, distances between sites are measured in meters, and the closest pairs of sites are approximately 7.2 m apart, so this prior favors a negligible correlation of only 0.0068 between the closest sites.

4 Bayesian inference for the spatial ZIP model

Since the likelihood (3) involves a high-dimensional integral, direct Bayesian inference for the spatial ZIP model is intractable even when using Monte Carlo methods. Therefore, we shall apply a data augmentation method similar to that proposed by Chib and Greenberg (1998) for their spatial probit model. Instead of drawing samples of the parameter θ from the posterior distribution $p(\theta|\mathbf{z}) \propto p(\mathbf{z}|\theta)\,\pi(\theta)$, samples of $(\theta, \mathbf{y})$ are drawn from $p(\theta, \mathbf{y}|\mathbf{z}) \propto p(\mathbf{z}|\theta, \mathbf{y})\,p(\mathbf{y}|\theta)\,\pi(\theta)$ using a Markov Chain Monte Carlo (MCMC) algorithm (Gilks et al. 1996). In our implementation, the elements of $\mathbf{y}, \alpha, \beta$, and γ are successively updated during each iterate of the MCMC algorithm. A hybrid algorithm is adopted, using Gibbs steps to update $\mathbf{y}$ and α, and Metropolis-Hastings steps to update β and γ.

4.1 MCMC algorithm

The conditional distribution of Y_i depends on the value of the data z_i. For $z_i = 0$,

$$\begin{aligned} p\left(y_i|\mathbf{y}_{(i)}, \mathbf{z}, \theta\right) &= p\left(y_i|\mathbf{y}_{(i)}, z_i = 0, \theta\right) \\ &\propto \left(2\pi\sigma^2_{\gamma,(i)}\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2_{\gamma,(i)}}\left(y_i - \mu_{(i)}\right)^2\right\} \\ &\quad \times \left\{I(y_i < 0) + I(y_i > 0)\exp\left\{-e^{\beta'\mathbf{x}_i}\right\}\right\}, \end{aligned} \tag{4}$$

where $\mu_{(i)} = \alpha'\mathbf{x}_{1i} + \mathbf{c}'_{\gamma,(i)}\Sigma^{-1}_{\gamma,(i)}\left(\mathbf{y}_{(i)} - \mathbf{X}_{(i)}\alpha\right)$, and $\sigma^2_{\gamma,(i)} = 1 - \mathbf{c}'_{\gamma,(i)}\Sigma^{-1}_{\gamma,(i)}\mathbf{c}_{\gamma,(i)}$. The matrix $\mathbf{X}_{(i)}$ is $\mathbf{X}_1$ with the ith row removed, $\Sigma_{\gamma,(i)}$ is Σ_γ with the ith row and ith column removed, and $\mathbf{c}_{\gamma,(i)}$ is the ith column of Σ_γ with the ith element removed. Note that the inverse of $\Sigma_{\gamma,(i)}$ can be readily obtained from Σ^{-1}_γ (Christensen et al. 1992). Sampling from (4) is straightforward. Let $q = \Phi\left(-\mu_{(i)}/\sigma_{\gamma,(i)}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution. With probability $p = q/\left(q + (1-q)\exp\left\{-e^{\beta'\mathbf{x}_i}\right\}\right)$ sample from $N\left(\mu_{(i)}, \sigma^2_{\gamma,(i)}\right)$ truncated to the right at zero, otherwise sample from $N\left(\mu_{(i)}, \sigma^2_{\gamma,(i)}\right)$ truncated to the left at zero. For $Z_i > 0$, the Y_i is conditionally distributed according to the truncated Gaussian distribution

$$\begin{aligned} p\left(y_i|\mathbf{y}_{(i)}, \mathbf{z}, \theta\right) &= p\left(y_i|\mathbf{y}_{(i)}, z_i > 0, \theta\right) \\ &\propto \left(2\pi\sigma^2_{\gamma,(i)}\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2_{\gamma,(i)}}\left(y_i - \mu_{(i)}\right)^2\right\} I(Y_i > 0). \end{aligned}$$

The algorithm selected for generating samples from the truncated Gaussian density function $\varphi(x)/\Phi(a)$; $x \le a$, depends on the truncation point a. For large a, a simple rejection sampling algorithm may be applied, selecting standard normal random variates X until a value $X \le a$ is attained. Here, the polar method of Box and Muller (1958) was used to generate the standard normal variates. Acceptance probabilities decrease with deceasing values of a. For small a, the rejection sampling algorithm of Marsaglia (1964) for generating variates from the tail of a Gaussian distribution yields higher acceptance probabilities. For Marsaglia's algorithm, acceptance probabilities increase with decreasing a. For the present application, Box and Muller's algorithm was used for $a > -0.28$, while Marsaglia's algorithm was used for $a \le -0.28$.

As determined on a Sun Blade 1000 computer, both algorithms generate approximately equal numbers of realizations per second when $a = -0.28$.

The conditional distribution of α given $\mathbf{y}, \mathbf{z}$, and γ is $N\left(\widetilde{\mu}_\alpha, \widetilde{\mathbf{V}}_\alpha\right)$,where

$$\widetilde{\mu}_\alpha = \widetilde{\mathbf{V}}_\alpha \left(\mathbf{V}_\alpha^{-1}\mu_\alpha + \mathbf{X}_1'\Sigma_\gamma^{-1}\mathbf{y}\right)$$

and

$$\widetilde{\mathbf{V}}_\alpha = \left[\mathbf{X}_1'\left(g_\alpha^{-1}\mathbf{I}+\Sigma_\gamma^{-1}\right)\mathbf{X}_1\right]^{-1}.$$

Samples from this multivariate normal distribution can readily be obtained using the Cholesky decomposition method.

The conditional distribution of β is

$$\begin{aligned} p\left(\beta|\mathbf{y},\mathbf{z},\alpha,\sigma,\gamma\right) &= p\left(\beta|\mathbf{y},\mathbf{z},\sigma,\gamma\right) \\ &\propto \exp\left\{\sum_{i=1}^{n}\left(z_i\left(\beta'\mathbf{x}_{2i}\right) - e^{\beta'\mathbf{x}_i}\right) - \frac{1}{2}\left(\beta-\beta_0\right)'\mathbf{V}_\beta^{-1}\left(\beta-\beta_0\right)\right\} \end{aligned}$$

Although it is feasible to sample directly from $p\left(\beta|\mathbf{y},\mathbf{z},\sigma,\gamma\right)$ using a rejection sampling algorithm, the rejection rate is unacceptably high. Therefore, estimates of β were updated using a Metropolis-Hastings step. Slow convergence was achieved under block updating of β, so the elements of β were updated separately. Given current values for $\mathbf{y}, \beta, \sigma, \gamma$, a candidate value β_j^* for the jth element of β is selected from $N\left(\beta_j, \psi_j^2 v_j\right)$, where ψ_j^2 is a tuning constant and v_j is the prior variance of β_j. The candidate value β_j^* is accepted with probability

$$\min\left\{\frac{p\left(\beta^*|\mathbf{y},\mathbf{z},\sigma,\gamma\right)}{p\left(\beta|\mathbf{y},\mathbf{z},\sigma,\gamma\right)}, 1\right\},$$

where β^* is obtained by replacing the jth element of β with β^*. The tuning constant ψ_j^2 controls the acceptance rate of the algorithm. If ψ_j^2 is too small, then the acceptance rate is high, but jump sizes are correspondingly small, yielding slow convergence. Conversely, the selection of high values for ψ_y^2 leads to larger jump sizes, but at the cost of lower acceptance rates. Following the recommendation of Gelman et al. (1996), ψ_j^2 shall be selected so as to yield empirical acceptance rates around 0.25. Improved performance of the algorithm was obtained by performing 12 subiterations of the Metropolis-Hastings step for the elements of β during each iteration of the MCMC algorithm (Carlin and Louis 2000, p. 160).

Finally, the conditional distribution of γ is

$$\begin{aligned} p\left(\gamma|\mathbf{y},\mathbf{z},\alpha,\beta,\sigma\right) &= p\left(\gamma|\mathbf{y},\alpha\right) \\ &\propto \left|\Sigma_\gamma\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}_1\alpha\right)'\Sigma_\gamma^{-1}\left(\mathbf{y}-\mathbf{X}_1\alpha\right)\right\}\pi\left(\gamma\right). \end{aligned}$$

Since it is difficult to sample directly from this distribution, values of γ shall be generated using the Metropolis-Hastings algorithm. Given current values of $\mathbf{y}, \alpha$, and γ, a candidate value γ^* for γ is generated using the method suggested by De Oliveira (2000): Take the logit transformation $\eta = \text{logit}\left(\gamma\right)$, and generate η^* from $N\left(\eta, \psi_\gamma^2\right)$, where the tuning parameter ψ_γ^2 is selected to achieve an empirical acceptance rate around 0.25 (Gelman et al. 1996). Here, the acceptance probability becomes

$$\min\left\{\frac{p\left(\gamma^*|\mathbf{y},\alpha\right)\gamma^*\left(1-\gamma^*\right)}{p\left(\gamma|\mathbf{y},\alpha\right)\gamma\left(1-\gamma\right)},1\right\}.$$

Again 12 subiterations of the this Metropolis-Hastings step were carried out for each iteration of the MCMC algorithm.

To achieve the desired acceptance rates, an adaptive algorithm similar to that proposed by Browne et al. (2002) was applied to select the tuning constants ψ required for the Metropolis-Hastings steps of the MCMC algorithm. The tuning constant ψ is set to an arbitrary starting value ($\psi_j = 0.5; j = 1,\ldots,p_2$ and $\psi_\gamma = 0.11$ in the present application), and the algorithm is run in batches of 100 iterations. The objective is to achieve an acceptance rate within a specified tolerance interval $(r-\Delta, r+\Delta)$. Following each batch of iterations, the empirical acceptance rate r^* for that batch is compared to tolerance interval. If $r^* > r+\Delta$, replace ψ with $\psi\left(2-(1-r^*)/(1-r)\right)^{-1}$; if $r^* < r-\Delta$, replace ψ with $\psi\left(2-r^*/r\right)$; and retain the current value of ψ if r^* falls inside the tolerance region. The adaptive procedure ends when 5 successive values of r^* fall within the tolerance region.

4.2 Diagnostics

The application of the above rules yields a Markov chain $\left\{\left(\theta^{(t)},\mathbf{y}^{(t)}\right): t=1,2,\ldots\right\}$, which following a sufficiently long burn-in period will be approximately distributed as $p\left(\theta,\mathbf{y}|\mathbf{z}\right)$ (Roberts and Smith 1993). We shall use the test for stationarity proposed by Heidelberger and Welch (1983), implemented in the CODA software package (Best et al. 1995) on R, to assess convergence of the Markov chain. Based on Schruben's (1982) Brownian bridge model, Heidelberger and Welch test for initial transients in the simulated chain. If an initial transient is detected, the test is repeated after discarding an initial percentage (10% by default) of the iterations. Additional iterations are discarded as necessary until a non-significant result can be reported. Thus, a recommended burn-in period is given for each parameter.

4.3 Point estimation and uncertainty assessment

Following a sufficient burn-in period b, a point estimator for the parameters θ can be obtained from the mean

$$\widehat{\theta} = \frac{1}{T-b}\sum_{t=b+1}^{T}\theta^{(t)}.$$

Here, b is taken to be the largest burn-in period reported by Heidelberger and Welch's (1983) procedure among all model parameters and parallel chains. Since successive values of $\theta^{(t)}$ are positively correlated, the sample variance–covariance matrix would yield an underestimate of the uncertainty of θ. Kass et al. (1998) suggest an approach based on the computation of effective sample size using estimates of the autocorrelation. Estimation of the autocorrelation requires a point estimator for θ. The use of $\widehat{\theta}$ or any other point estimator introduces a small bias in the estimator for the autocorrelation function, and hence a bias in the resulting variance estimator. Improved performance can be achieved through variography. Define the multivariate semivariogram

$$\Gamma\left(h\right) = \frac{1}{2}E\left\{\left(\theta^{(t)} - \theta^{(t+h)}\right)\left(\theta^{(t)} - \theta^{(t+h)}\right)'\right\}.$$

Note that $\Gamma\left(h\right) \to \text{var}\left(\theta|\mathbf{z}\right)$ as $h \to \infty$. The method of moments estimator

$$\widehat{\Gamma}\left(h\right) = \frac{1}{2\left(T-b-h\right)}\sum_{t=b+1}^{T-h}\left(\theta^{(t)} - \theta^{(t+h)}\right)\left(\theta^{(t)} - \theta^{(t+h)}\right)'$$

is an unbiased estimator for $\Gamma\left(h\right)$, and under suitable fourth-moment properties for the Markov chain $\left\{\theta^{(t)}\right\}$, $\widehat{\Gamma}\left(h\right)$ converges in probability to $\Gamma\left(h\right)$ as $T \to \infty$. The elements of $\widehat{\Gamma}\left(h\right)$ can be plotted against lag h to determine the lag ℓ at which all elements are close to their respective asymptotes. The failure of one or more elements of $\widehat{\Gamma}\left(h\right)$ to converge towards an asymptote provides an additional diagnostic indicating whether the Markov chain has yet to converge in distribution to the posterior. Once ℓ has been determine, the posterior variance may be estimated by

$$\widehat{\text{var}}\left(\theta|\mathbf{z}\right) = \frac{1}{M}\sum_{h=\ell}^{T-b-1}\sum_{t=b+1}^{T-h}\left(\theta^{(t)} - \theta^{(t+h)}\right)\left(\theta^{(t)} - \theta^{(t+h)}\right)', \tag{5}$$

where M is the number of elements in the double sum.

5 Implementation on oak regeneration data

The Pennsylvania oak regeneration data were available for 1,331 plots spread over 38 mixed-oak forest stands (Fig. 1). To improve computational efficiency, and to reduce floating-point errors in manipulations of large variance–covariance matrices, data from different forest stands were assumed to be independent. This reduced the largest matrix to be manipulated to 39 × 39, corresponding to the size of the largest stand. Treating the stands as independent samples is justified since the estimated range of spatial correlation for all three species falls well below the distances separating the closest stands (Table 3).

The zero-inflated Poisson regression model has two component parts, a spatial probit model for the presence/absence of oak seedlings, and a Poisson regression model for seedling counts. When we attempted to include the same spatial covariates in both model components, the MCMC algorithm performed poorly, showing little evidence for convergence of model parameters to the posterior distribution. This outcome is likely the consequence of high posterior correlations among model parameters. In the following, the physical variables, slope shape (shape), slope percentage (slope), slope aspect (aspect), exposure angle (exposure), and elevation were used to model the presence/absence of oak seedlings, while the biotic variables, counts of adult trees of the same species, and percent covers of heather shrubs and hayscented fern were used to model the abundance of oak seedlings where they were deemed to be present. According to Shelford's (1913) theory of tolerance, each and every plant species is able to exist and reproduce successfully only within a definite range of environmental conditions. The conditions under which a species can exist in isolation are its potential range, and physical variables can serve as good measurements of their potential range. Within the species potential range, seed availability, competition, mutualism, and other biotic factors can affect the abundance of species. Aspect θ is a circular variable that measures the direction of the steepest descent. Attention must be give to its periodic nature during model building. In the following, the effect of aspect will be modeled through a linear combination of the two variables $\sin\theta$ and $\cos\theta$.

5.1 Tuning constants

Implementation of Bayesian inference for the zero-inflated model for species counts requires the specification of starting values for the tuning constants of the Metropolis-Hastings steps of the algorithm. Optimal tuning constants are increasing functions of the posterior variances of their respective parameters, which in turn are decreasing functions of sample size. The tuning constants ψ_j for the Poisson regression coefficients β_j were initialized at 0.5, and the tuning constant ψ_γ for the spatial correlation parameter γ was initialized at 0.11. For each species, samples from the posterior distribution were obtained from 40,000 iterates of the proposed MCMC algorithm. The adaptive algorithm for tuning constant specification converged within 1,200 iterates for chestnut oak and white oak, and within 1,700 iterates for red oak. These initial iterates will be removed in all of the following analyses. Table 1 presents the final tuning constants and acceptance rates for the parameters of the Poisson part of the model, and the spatial dependence parameter γ. Acceptance rates fall between 0.206 and 0.277, all close to the optimal rate of 0.25 suggest by Gelman et al. (1996).

5.2 MCMC diagnostics

In Fig. 3, parameter values are plotted against iterate number for chestnut oak seedlings for every tenth iterate. In the interest of saving space, only selected parameters were plotted. The remaining parameters behave similarly to those of the first six parameters in the plot, and similar results were obtained for the other two oak species. The range of spatial correlation is $r = -3/\log\gamma$; pairs of sites further than r apart are negligibly correlated. Since the first 1,000 iterates of each parameter vary quite widely, these iterates are not included in the plots. The results suggest that the regression coefficients for both the probit and Poisson components of the model show excellent mixing properties. However, coefficients for the Poisson part of the model converge more slowly than the coefficients for the probit part of the model. The range shows evidence of stronger temporal dependence across iterates than the remaining parameters of the model. All variables for all three species passed the Heidelberger and Welch test for convergence, in some cases after some initial portion of the iterates

Table 1 Tuning constants and acceptance rates for the Metropolis-Hastings steps of the MCMC algorithm

Parameter		Species					
	Initial	Chestnut oak		White oak		Red oak	
	Tuning	Tuning	Acceptance	Tuning	Acceptance	Tuning	Acceptance
Intercept	0.50	0.406	0.238	0.269	0.236	2.619	0.221
Overstory	0.50	0.312	0.227	0.086	0.246	3.244	0.206
Low Shrubs	0.50	0.016	0.236	0.012	0.215	0.117	0.239
Ferns	0.50	6.096	0.277	6.274	0.270	6.434	0.260
γ	0.11	0.227	0.265	0.219	0.262	0.237	0.261

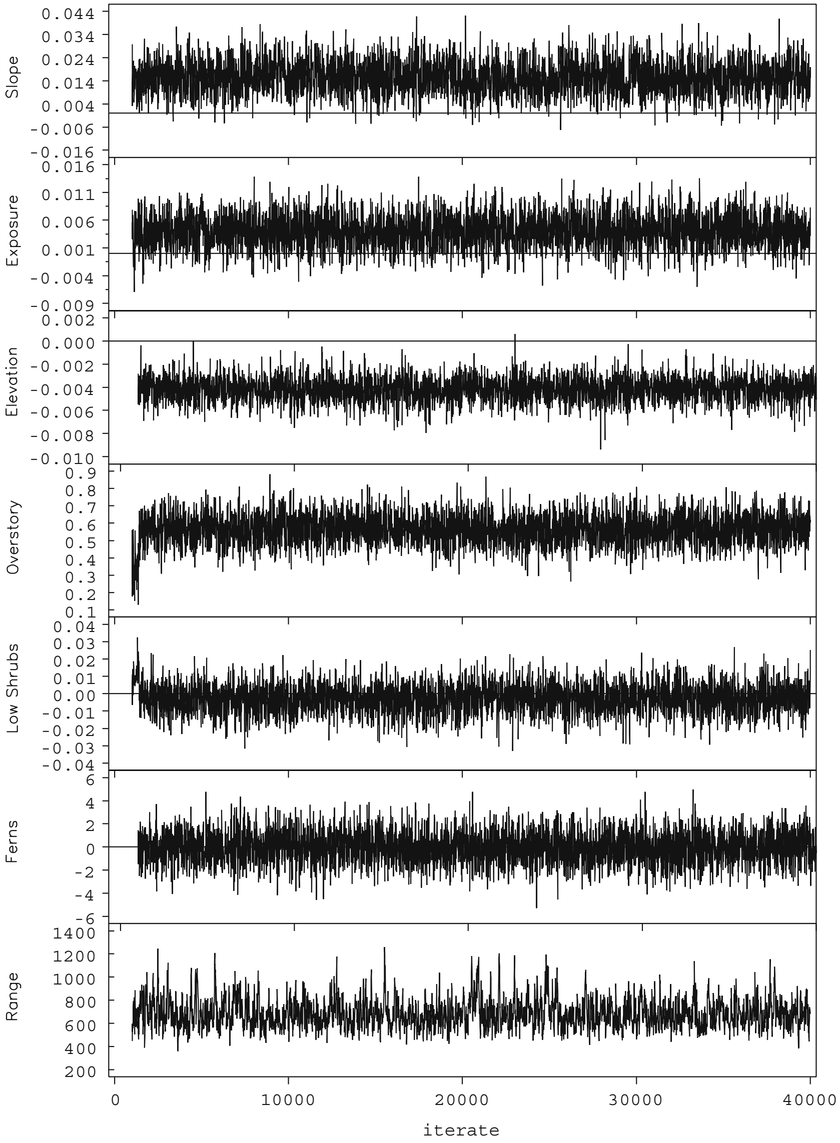

Fig. 3 Iterates of the MCMC algorithm for selected parameters of the spatial zero-inflated Poisson model for chestnut oak seedlings counts

were removed from the analysis as per standard procedure with this diagnostic. In the remaining analyses, the burn-in period for each species will be set to the maximum starting iteration required to pass the Heidelberger and Welch test. These burn-periods are 5,861, 14,401, and 1,701, respectively, for chestnut, white, and red oak.

Figure 4 presents probability density functions estimated from the samples from the posterior distribution for chestnut oak. Again, only selected variables were plotted. The parameters for both the probit and Poisson portions of the model tend to be approximately symmetrically distributed for this species and the other two oak

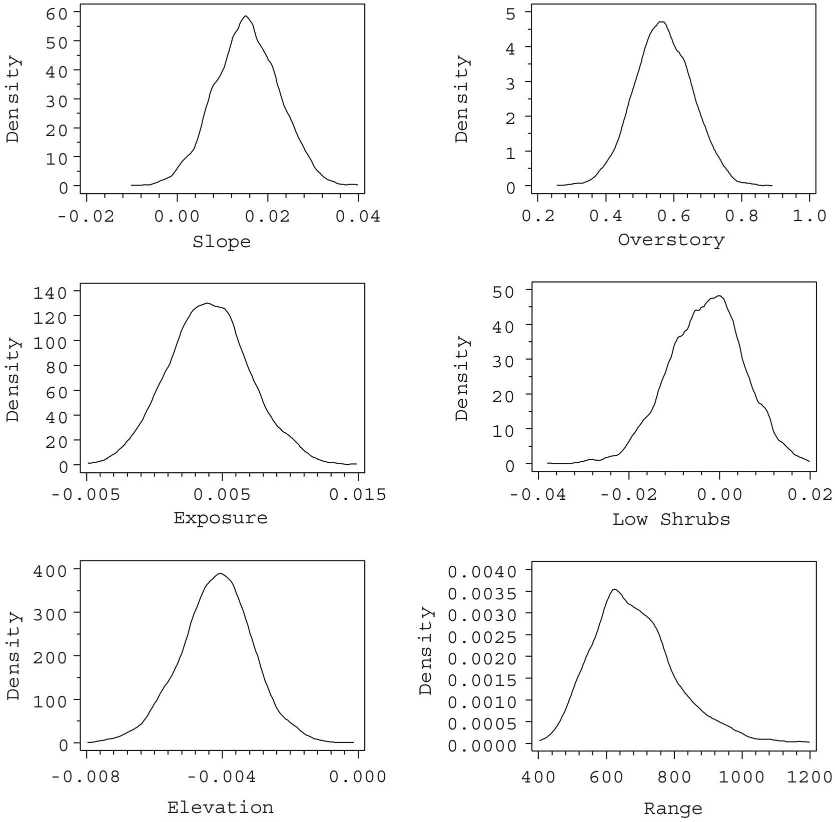

Fig. 4 Kernel density estimates of the posterior distribution of select model parameters for chestnut oak

species. In contrast, the range parameter shows a moderately skewed distribution for all three species.

Semivariograms are plotted against lag iterate difference for the Monte Carlo samples of select variables for chestnut oak are presented in Fig. 5. This figure shows that the regression coefficients for both the probit and Poisson parts of the model show very small ranges of correlation across iterates. The range of spatial correlation, however, has a long range of correlation across iterates. Only after the lag exceeds 84 iterates does the correlation between successive iterates falls below 0.05. Therefore, we take $\ell = 84$ in expression (5) for the estimated posterior variance of each parameter.

5.3 Effect of prior choice

To investigate the effect of prior choice on Bayesian inference for ZIP model parameters, four different prior models were considered (Table 2). Two levels of g were considered for the Zellner g-prior for the regression coefficients of the probit and Poisson parts of the model. Three levels of the hyperparameters a and b were considered for the Beta(a, b) prior for the spatial correlation parameter γ. The posterior means and standard deviations of the parameters for chestnut oak are depicted in

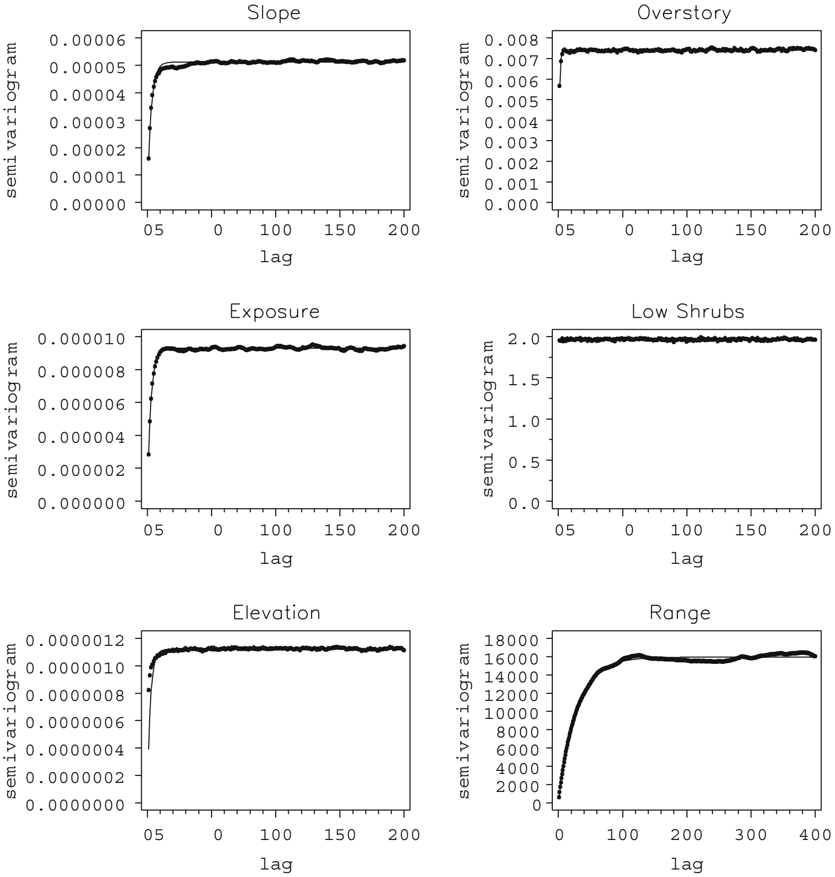

Fig. 5 Semivariograms against lag itereate difference for posterior samples of select model parmeters for chestnut oak

Table 2 Prior models

Prior model	Zellner g-prior for α, β	Beta prior for γ	
	g_α, g_β	a	b
1	1,000	1	1
2	500	1	1
3	500	3/2	1
4	500	1	3/2

Fig. 6; similar results were obtained for the remaining two oak species. This figure indicates that prior model choice had little effect on the posterior means and standard deviations of the model parameters.

5.4 Results

The estimated ranges of spatial correlation do not differ greatly among the three oak species (Table 3). They range between 684 m for chestnut oak to 894 m for white oak.

Table 4 shows the Bayes estimates of the regression coefficients together with their posterior standard errors of the probit part of the model. For chestnut oak

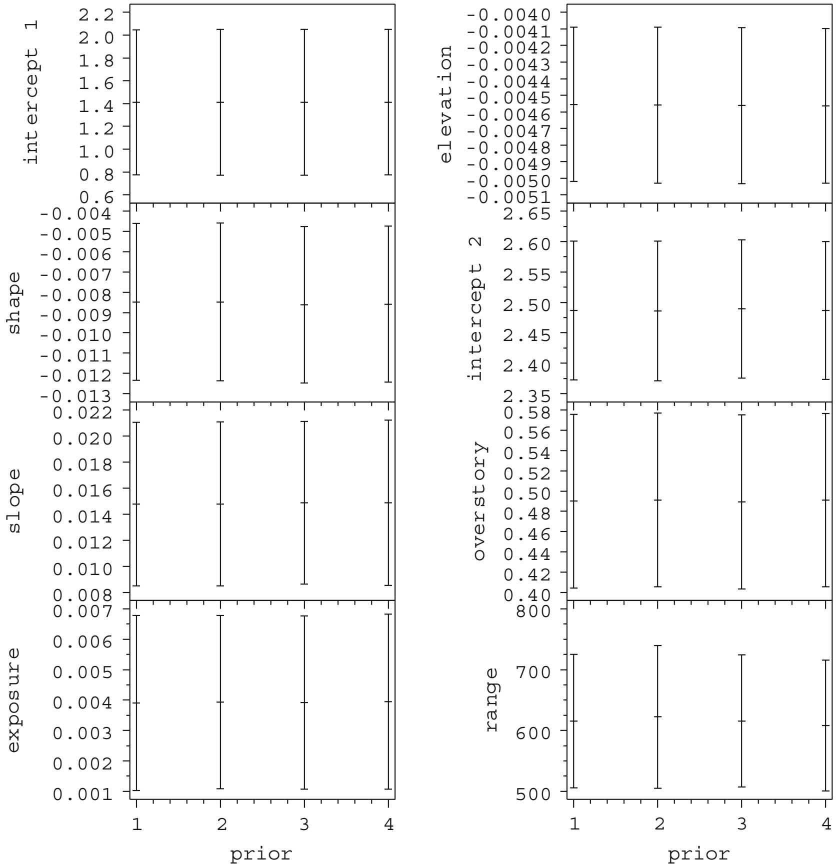

Fig. 6 Posterior means plus and minus their standard errors for various prior models for chestnut oak model parameters

Table 3 Estimated range of spatial correlation

Species	Range	Standard error
Chestnut oak	684.0	123.6
White oak	894.4	181.5
Red oak	743.2	160.9

regeneration, slope percent and elevation have significant influence on its range of distribution. Chestnut oak favors regions with steeper slopes and low elevation. McQuilkin (1990) also reported that chestnut oak is most commonly found on steeper slopes with shallow soils. Chestnut oak has relatively higher abundance in the relative low elevation Ridge and Valley than the Appalachian Plateau physiographic provinces. The only significant factor for white oak is elevation, which agrees with the former study that white oak grows best on lower slopes and coves (Rogers 1990). No factors are found to have significant influences on the distribution of red oak regeneration.

For the Poisson part of the model, the Bayes estimates of the regression coefficients and their posterior standard errors are given in Table 5. Among the three biotic

Table 4 Inference for the probit part of the model

Variable	Species		
	Chestnut oak	White oak	Red oak
Intercept	1.24489	1.47971	1.00145
	(0.66229)	(0.66784)	(1.01082)
Shape	−0.00863	0.00529	0.00420
	(0.00473)	(0.00476)	(0.00857)
Slope	0.01541	−0.01213	−0.01000
	(0.00721)	(0.00728)	(0.01157)
Exposure	0.00404	−0.00122	0.00385
	(0.00306)	(0.00286)	(0.00528)
Elevation	−0.00416	−0.00348	−0.00025
	(0.00106)	(0.00089)	(0.00132)
$\sin\theta$	−0.06025	0.00620	0.04533
	(0.07982)	(0.07832)	(0.12984)
$\cos\theta$	−0.04879	−0.02539	0.05578
	(0.07208)	(0.07269)	(0.12163)

Table 5 Inference for the Poisson part of the model

Variable	Species		
	Chestnut oak	White oak	Red oak
Intercept	2.41783	3.37363	−1.61527
	(0.21747)	(0.12318)	(1.77110)
Overstory	0.56783	−0.17890	1.10645
	(0.08570)	(0.08317)	(0.61342)
Low shrubs	−0.00307	0.00155	−0.01280
	(0.00838)	(0.00544)	(0.07918)
Haysented ferns	−0.04947	−0.06181	−0.06433
	(1.40565)	(1.41643)	(1.39812)

factors, overstory tree density is the only factor that is significant or marginal significant for all the three oak species. Not surprisingly, chestnut oak and red oak is strongly favored by the presence of the same species in the canopy above the plot. However, the negative association between overstory and understory white oak abundance is puzzling. After reviewing the original data, we found that most of the high density white oak plots have small stem diameter, which might indicate that these plots are relatively young and have not reach the cycle of regeneration stage yet.

Acknowledgements We would like to thank an anonymous referee for his helpful comments. This work was partially supported by National Science Foundation Grant DMS-0322673. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Abramowitz M, Stegun IA (1965) Handbook of mathematical functions. Dover, New York

Agarwal DK, Gelfand AE, Citron-Pousty S (2002) Zero-inflated models with application to spatial count data. Environ Ecol Stat 9:341–355

Ashford JR, Sowden RR (1970) Multivariate probit analysis. Biometrics 26:535–546

Best NG, Cowles MK, Vines SK (1995) CODA Manual version 0.30. MRC Biostatistics Unit, Cambridge, UK

Box GEP, Muller ME (1958) A note on the generation of random normal deviates. Ann. Math. Stat 29:610–611

Browne WJ, Draper D, Goldstein H, Rasbash J (2002) Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. Compu Stat Data Anal 39:203–225

Carlin BP, Louis TA (2000) Bayes and empirical bayes methods for data analysis. Chapman and Hall, Boca Raton, FL

Chib S, Greenberg E (1998) Analysis of multivariate probit models. Biometrika 85:347–361

Christensen R, Johnson W, Pearson LM (1992) Prediction diagnostics for spatial linear models. Biometrika 79:583–591

Cressie N (2001) Statistics for spatial data. Wiley, New York

Crow TR (1988) Reproduction mode and mechanisms for self-replacement of northern red oak (*Quercus rubra*) – a review. Forest Sci 34:19–40

De Oliveira V (2000) Bayesian prediction of clipped Gaussian random fields. Comput Stat Data Anal 34:99–314

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics. Appl Stat 47:299–350

Gelman A, Roberts GO, Gilks WR (1996) Efficient Metropolis jumping rules. In: Bayesian statistics 5. Bernardo JM, Berger JO, Dawid AP, and Smith AFM (eds) Oxford University Press, Oxford, pp 599–607

Gilks WR, Richardson S, Spiegelhalter DJ (1996), Markov Chain Monte Carlo in Practice, Chapman and Hall, London

Gotway CA, Stroup WW (1997) A generalized linear model approach to spatial data analysis and prediction. J Agric Biol Environ Stat 2:157–178

Gotway CA, Wolfinger RD (2003) Spatial prediction of counts and rates. Stat. Med 22:1415–1432

Hall DB (2000) Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 56:1030–1039

Heagerty PJ, Lele SR (1998) A composite likelihood approach to binary spatial data. J Am Stat Assoc 93:1099–1111

Heidelberger P, Welch PD (1983) Simulation run length control in the presence of an initial transient. Oper Res 31:1109–1144

Kass RE, Carlin BP, Gelman A, Neal R (1998) Markov chain Monte Carlo in practice: A roundtable discussion. Am Stat 52:93–100

Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34:1–14

Leathwick JR, Austin MP (2001) Competitive interactions between tree species in New Zealand's old-growth indigenous forests. Ecology 82:2560–2573

Lorimer CG (1992) Causes of the oak regeneration problem. In: Loftis D, McGee Ce (eds) Oak regeneration: serious problems, practical recommendations. USDA Forest Service, Southeastern Forest Experiment Station. General Technical Report SE-84

Marsaglia G (1964) Generating a variable from the tail of the normal distribution. Technometrics 6:101–102

Matern B (1960) Spatial variation. Meddelanden fran Statens Skogsforskningsinstitut, 49, No. 5. Alamaenna Foerlaget, Stokholm.

McQuilkin R (1990) In: Silvics in North America. USDA Forest Service. Agriculture handbook, Vol 654, pp 605–613

O'Neill MF, Faddy MJ (2003) Use of binary and truncated negative binomial modelling in the analysis of recreational catch data. Fish Res 60:471–477

Roberts GO, Smith AFM (1993) Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stoch Process Appl 49, 207–216

Rogers R (1990) In: Silvics in North America. USDA Forest Service. Agriculture handbook, Vol 654, 605–613

Schruben LW (1982) Detecting intialization bias in simulation experiments. Oper Res 30:569–590

Shelford V (1913) Animal communities in temperate America. University of Chicago Press, Chicago

Stein ML (1999) Interpolation of spatial data. Springer, New York

Weir IS, Pettitt AN (1999). Spatial modelling for binary data using a hidden conditional autoregressive Gaussian process: a multivariate extension of the probit model. Stat Comput 9:77–86

Weir IS, Pettitt AN (2000) Binary probability maps using a hidden conditional autoregressive Gaussian process with an application to Finnish common toad data. Appl Stat 49:473–484

Welsh AH, Cunningham RD, Donelly CF, Lindenmayer DB (1996), Modelling the abundance of a rare species: Statistical modes for counts with extra zeros. Ecol Model 88:297–308

Welsh AH, Cunningham RD, and Chambers RL (2000) Methodology for estimating the abundance of rare animals: seabird nesting on North East Herald Cay. Biometrics 56:22–30

Van Iersel MW, Oetting RD, Hall DB (2000). Imidacloprid applications by subirrigation for control of silverleaf whitefly (Homoptera: Aleyrodidae) on poinsettia. J Econ Entomol 93:813–819

Van Iersel MW, Oetting RD, Hall DB, Kang JG, (2001) Application technique and irrigation method affect imidacloprid control of silverleaf whiteflies (Homoptera: Aleyrodidae) on poinsettias. J Econ Entomol 94:666–672

Vieira AMC, Hinde JP, Demetrio CGB (2000) Zero-inflated proportion data models applied to a biological control assay. J Appl Stat 27:373–389

Zellner A (1986) On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In: Goel P, Zellner A (eds) bayesian inference and decision techniques: essays in honor of Bruno de Finetti. North Holland Publishing Company, New York, pp 233–243

Biographical Sketches

Stephen L. Rathbun is Associate Professor of Biostatistics in the Department of Health Administration, Biostatistics and Epidemiology at the University of Georgia.

Songlin Fei is Assistant Professor of Forestry in the Department of Forestry at the University of Kentucky.