Collection of Biostatistics Research Archive

COBRA Preprint Series

Year 2008

Paper 47

Joint Spatial Modeling of Recurrent Infection and Growth with Processes Under Intermittent Observation

Farouk S. Nathoo*

*University of Victoria, nathoo@math.uvic.ca

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/cobra/art47>

Copyright ©2008 by the author.

Joint Spatial Modeling of Recurrent Infection and Growth with Processes Under Intermittent Observation

Farouk S. Nathoo

Abstract

In this article we present new statistical methodology for longitudinal studies in forestry where trees are subject to recurrent infection and the hazard of infection depends on tree growth over time. Understanding the nature of this dependence has important implications for reforestation and breeding programs. Challenges arise for statistical analysis in this setting with sampling schemes leading to panel data, exhibiting dynamic spatial variability, and incomplete covariate histories for hazard regression. In addition, data are collected at a large number of locations which poses computational difficulties for spatiotemporal modeling. A joint model for infection and growth is developed; wherein, a mixed non-homogeneous Poisson process, governing recurring infection, is linked with a spatially dynamic nonlinear model representing the underlying height growth trajectories. These trajectories are based on the von Bertalanffy growth model and a spatially-varying parametrization is employed. Spatial variability in growth parameters is modeled through a multivariate spatial process derived through kernel convolution. Inference is conducted in a Bayesian framework with implementation based on hybrid Monte Carlo. Our methodology is applied for analysis in an eleven year study of recurrent weevil infestation of white spruce in British Columbia.

1. Introduction

The white pine weevil is a native North American insect that poses a significant threat to spruce and pine trees, with consequences of infestation felt on both economic and environmental scales. As such, there has recently been an enormous investment on studying this disease and its relationship with other ecological processes. In particular, understanding the relationship between weevil infection and tree growth has important implications in directing selections for seed orchards, as well as in breeding program development. To this end, longitudinal studies of disease ecology often collect data on recurrent infection and plantation development. Challenges arise for statistical analysis in this setting with observation and sampling schemes leading to interval censoring and panel data exhibiting dynamic spatial variability. In addition, data are collected at a large number of locations which poses computational difficulties for spatiotemporal modeling.

In this article we present new statistical methodology for such studies, where the hazard of infection is modeled jointly with tree growth over time. A joint model for infection and growth is developed; wherein, a mixed non-homogeneous Poisson process, governing recurring infection, is linked with a nonlinear mixed model representing the underlying height growth trajectories. These trajectories are based on the von Bertalanffy growth model, and a spatially-varying parametrization is employed. Spatial variability in growth parameters is modeled through a multivariate spatial process derived through kernel convolution. Inference is conducted within a hierarchical Bayes framework, with implementation based on hybrid Monte Carlo. Our methodological developments are motivated by an eleven year study of pine weevil infestation, where information relating to infection and growth was obtained for analysis. The plantation under study is depicted in Figure 1, marking the locations of $n = 4330$ interior spruce trees.

Recent work developing spatiotemporal regression models of forest insect outbreak has focused on methods for correlated binary data. Adopting a Markov random field approach, Zhu et al. (2005) and Zhu et al. (2008) develop space-time autologistic models for prediction of pine beetle infestation. There, spatial and temporal dependence is introduced through a neighborhood structure, defining conditional distributions over a lattice. Model fitting and inference is based on pseudo-likelihood and Monte Carlo maximum likelihood methods. An alternative approach, considered by Nathoo and Dean (2007), develops a two-state transitional model for describing pine weevil infection, with disease immunity modeled through a mover-stayer structure. Spatially correlated random effects are introduced at the second level of a hierarchical model, and inference proceeds through a Monte Carlo EM algorithm. These methods are appropriate when data on infection arise through a binary response, collected in discrete-time, with time-dependent covariates that are completely observed. In contrast, the data in our experiment arise through a more complicated structure, in the form of recurrent events with interval censoring, and missing information on time-dependent covariates.

A primary focus of our analysis relates to how the hazard of infection depends on the rate of height growth over time. A second objective lies in studying height growth itself, and in characterizing spatial variability in properties of growth. Difficulty arises through the sampling of the growth process, which leads to an incomplete covariate history for hazard regression. To address our objectives within a unified setting, we develop our models and analysis within a framework allowing for joint modeling of recurrent events and longitudinal growth data (Ibrahim, Chen and Sinha 2001; Tsiatis and Davidian 2004).

Hierarchical Bayes methods for joint modeling with interval censored recurrent events have been developed for analysis in epidemiologic studies of tumor occurrence. Examples of such work include Sinha and Maiti (2004), who develop a model for panel-counts under dependent termination, and Dunson et al. (2004), who develop a joint model for describing recurrent lesion onset and growth. These methods for joint modeling assume independence

between subjects. In our forest ecology context, interaction through local properties of the landscape leads to spatial variability. It is of considerable interest to describe this variability and so spatially explicit models are required. Most importantly, tree growth is subject to influence from several factors, including competition and micro-site variability, arising from spatial variation in soil, topographic, geologic and micro-meteorological factors (Fox et al., 2007). These factors create spatial dependence in tree growth measurements and; moreover, this spatial structure is expected to evolve over time as the plantation ages. Spatial modeling in our application is thus challenging, and this difficulty is compounded for modeling over a large set of spatial locations.

We model tree growth over time using a spatially dynamic nonlinear model. Here, location specific growth parameters are based on a multivariate spatial process derived through kernel convolution (Higdon, 1998; Higdon, 2002). Kernel convolution models, also known as *spatial moving averages*, provide a flexible and intuitive construction for a spatial process. Importantly, the construction is useful in the analysis of large spatial datasets when used in conjunction with dimension-reduced approximations. These approximations have recently been studied by Cressie and Pavlicová (2002) and Xia and Gelfand (2005), within the context of univariate spatial modeling. Multivariate spatial models based on kernel convolution have also been developed by Ver Hoef and Barry (1998) and further by Ver Hoef et al. (2004) for cokriging with Gaussian data. Inference there is conducted within a frequentist setting, with estimation based on restricted maximum likelihood. We shall adopt a similar specification for modeling spatially-varying growth parameters at the second level of our hierarchical model formulation.

Our primary methodological contribution is the development of a framework for joint modeling of recurrent infection and growth applicable in situations where the underlying trajectory function, representing growth, is *spatially-varying*. It is important to note that the methods proposed here are quite general, and could be applied usefully in other applications

of hazard regression where covariates are both intermittently observed and spatially-varying.

The modeling of trajectories in our setting is closely related to methods for spatially correlated functional data. Within this context, Baladandayuthapani et al. (2008) develop methods for hierarchical functional data based on penalized regression splines with spatially-varying coefficients. There, the regression splines allow for flexibility in modeling a functional response; however, the spatial modeling is based on a separable multivariate structure which can be overly restrictive. Banerjee and Johnson (2006) develop spatial growth curve models based on a linear specification. Spatial variability in the intercept and slope coefficients is modeled through a flexible bivariate spatial process, allowing for non-stationarity and multi-resolution spatial dependence. As a linear specification is not appropriate in our application, we instead opt for an ecologically well motivated mechanistic model, specified through the von Bertalanffy differential equation (Garcia 1983). While the aforementioned papers focus on functional or growth data as the primary, and only response, we are faced with the additional challenge of jointly modeling growth with recurring infection, and estimating the impact of growth on the hazard of infection over time. Our treatment of this modeling problem is fully Bayesian, with implementation based on Markov chain Monte Carlo (MCMC). Motivated by poor performance of MCMC samplers based on random walk updating, we develop an algorithm based on hybrid Monte Carlo (Gustafson, 1997; Ishwaran, 1999).

The remainder of this article is organized as follows. In Section 2 we specify our hierarchical joint spatial modeling formulation. Section 3 develops Bayesian inference for our model, and presents computational methods based on hybrid Monte Carlo. In Section 4 we present a spatiotemporal analysis of pine weevil infestation and tree growth in our motivating application. The goodness-of-fit of our model to various important aspects of the data is also taken up in this section. Section 5 concludes with discussion and comments on future work.

2. Joint Spatial Model for Recurrent Events and Height Growth

We develop our models within the context of our motivating application where, upon planting, trees are susceptible to recurring infection by pine weevil. Data relating to both infection and growth are observed for n trees each having spatial location $\mathbf{s}_i = (s_{x_i}, s_{y_i})' \in \mathbb{D} \subset \mathbb{R}^2$ (depicted in Figure 1). For the i^{th} tree, we represent recurring infection through a counting process $N_i(t)$ which denotes the total number of infections occurring in $(0, t]$. Data on the infection process arise through panel observation at a sequence of M_1 time points $t_1^{(N)}, \dots, t_{M_1}^{(N)}$ resulting in interval counts $N_{ij} = N_i(t_j^{(N)}) - N_i(t_{j-1}^{(N)})$, $i = 1, \dots, n$, $j = 1, \dots, M_1$, where N_{ij} represents the number of pine weevil attacks occurring in $(t_{j-1}^{(N)}, t_j^{(N)}]$ and $t_0^{(N)} = 0$. Here, we wish to investigate the dependence of the hazard function, governing the evolution of $N_i(t)$, to a second process $H_i(t)$ representing tree height, and sampled at a total of M_2 time points $t_1^{(H)}, \dots, t_{M_2}^{(H)}$.

We relate height growth to the hazard of infection through a joint model specified in a hierarchical Bayes framework. To model the process governing recurrent infection, we adopt a nonhomogeneous Poisson process incorporating a tree specific frailty. We model the height data as arising from a nonlinear mixed growth curve model. The two submodels representing growth and infection are then linked through an underlying spatial process. We specify a multivariate spatial process through kernel convolution, and discuss dimension reduction to facilitate computational implementation for large n .

2.1. Nonlinear Spatial Random effect Model for Height Growth

Our hierarchical specification assumes, at the first level, that the height measurements are drawn from a log-normal mixed model

$$Y_{ij} = \log H_i(t_j^{(H)}) = x_{\mathbf{s}_i}(t_j^{(H)}) + \epsilon_{ij}, \quad (1)$$

where $\epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$ accommodates measurement error, and $x_{\mathbf{s}_i}(t)$ is a *spatially-varying trajectory function* associated with the i^{th} tree at location $\mathbf{s}_i \in \mathbb{D}$. The trajectory is based on a nonlinear parametric growth curve, where we allow parameters to depend on location through a multivariate spatial process.

The Von Bertalanffy growth model is an established model for describing the growth of juvenile spruce trees (see for example, Garcia 1983). For the i^{th} tree, conditional on site specific parameters $b_1(\mathbf{s}_i) > 0$ and $b_2(\mathbf{s}_i) > 0$, the trajectory is specified through a first order differential equation

$$\frac{dx_{\mathbf{s}_i}(t)}{dt} = b_2(\mathbf{s}_i)[b_1(\mathbf{s}_i) - x_{\mathbf{s}_i}(t)], \quad i = 1, \dots, n. \quad (2)$$

Here, $b_1(\mathbf{s}_i)$ represents the final or asymptotic height (on log scale), and $b_2(\mathbf{s}_i)$ relates the rate of growth at time t , as proportional to the difference between the final and current height at time t . Integrating (2) yields the growth function

$$x_{\mathbf{s}_i}(t) = b_1(\mathbf{s}_i)[1 - \exp\{-b_2(\mathbf{s}_i)(b_3(\mathbf{s}_i) + t)\}], \quad (3)$$

where $b_3(\mathbf{s}_i) > 0$ is an additional parameter arising through a constant of integration. Spatial modeling of the growth parameters $\mathbf{b}(\mathbf{s}_i) = \{b_1(\mathbf{s}_i), b_2(\mathbf{s}_i), b_3(\mathbf{s}_i)\}'$ is discussed in Section 2.3.

2.2. Nonhomogeneous Mixed Poisson Process Model for Recurrent Infection

Conditioning on the history $x_{\mathbf{s}_i}^{(H)}(t) = \{x_{\mathbf{s}_i}(u), 0 \leq u < t\}$, we adopt a hierarchical nonhomogeneous Poisson process specification for $N_i(t)$, incorporating the trajectory as a spatially-varying covariate in a time-dependent Cox model

$$\lambda_i\{t|x_{\mathbf{s}_i}^{(H)}(t)\} = \lambda_0(t)\omega_i \exp\{\beta x_{\mathbf{s}_i}(t)\}. \quad (4)$$

The baseline intensity $\lambda_0(t)$ is assumed piecewise constant between inspection intervals $\lambda_0(t) = \lambda_j$ for $t \in (t_{j-1}^{(N)}, t_j^{(N)}]$. In (4), β characterizes the dependence of the infection process on height growth, and ω_i is a tree specific frailty, for which we assume $\omega_i \stackrel{iid}{\sim} \text{Gamma}(\eta, \eta)$, with $\eta^{-1} = \text{var}[\omega_i] > 0$ quantifying the degree of extra-Poisson variation in the interval counts. A specification incorporating spatial correlation between frailties is also possible (see for example, Li and Ryan, 2002; Banerjee et al. 2003) and would allow for spatial smoothing of residual heterogeneity. Note that a prime focus here is on investigating relationships with height, and spatial properties of height growth, so we direct attention to spatial modeling of the trajectory functions.

2.3. Spatial Model for Growth Parameters

To allow for spatial variability in the growth parameters $b_1(\mathbf{s}_i)$, $b_2(\mathbf{s}_i)$, $b_3(\mathbf{s}_i)$, we shall assume these arise through a partial realization of a trivariate spatial process $\mathbf{b}(\mathbf{s}) = \{b_1(\mathbf{s}), b_2(\mathbf{s}), b_3(\mathbf{s})\}'$, $\mathbf{s} \in \mathbb{D}$. Let $z_l(\mathbf{s}) = \log\{b_l(\mathbf{s})\}$, $l = 1, 2, 3$, and define the stationary Gaussian spatial process $\mathbf{z}(\mathbf{s}) = \{z_1(\mathbf{s}), z_2(\mathbf{s}), z_3(\mathbf{s})\}'$ through kernel convolution

$$\mathbf{z}(\mathbf{s}) = \boldsymbol{\mu}_z + \int_{\mathbb{R}^2} \mathbf{K}(\mathbf{u} - \mathbf{s}) \mathbf{T} \mathbf{w}(\mathbf{u}) d\mathbf{u}, \quad \mathbf{s} \in \mathbb{D}, \quad (5)$$

where $\boldsymbol{\mu}_z = (\mu_{z_1}, \mu_{z_2}, \mu_{z_3})'$ is the process mean; $\mathbf{K}(\mathbf{u}) = \text{Diag}\{K_1(\mathbf{u}), K_2(\mathbf{u}), K_3(\mathbf{u})\}$ where $K_l(\mathbf{u})$, $l = 1, 2, 3$, is a positive valued smoothing kernel; $\mathbf{w}(\mathbf{u}) = \{w_0(\mathbf{u}), \dots, w_3(\mathbf{u})\}'$ where $w_k(\mathbf{u})$, $k = 0, 1, 2, 3$, is a mean zero continuous white noise process on $\mathbb{R}^2$, with $w_k(\mathbf{u})$ independent of $w_m(\mathbf{u})$ when $k \neq m$; and $\mathbf{T}$ is 3×4 matrix, depending on parameters

$\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)'$ with $|\rho_l| < 1$, and taking the form

$$\mathbf{T}_{km} = \begin{cases} \rho_k & \text{if } m = 1 \\ \sqrt{1 - \rho_k^2} & \text{if } m = k + 1 \\ 0 & \text{o.w.} \end{cases}$$

The covariance functions resulting from the construction (5) are given by

$$c_{ll}(\mathbf{h}) = \text{cov}\{z_l(\mathbf{s}), z_l(\mathbf{s} + \mathbf{h})\} = \int_{\mathbb{R}^2} K_l(\mathbf{u})K_l(\mathbf{u} - \mathbf{h})\mathbf{d}\mathbf{u}, \quad l = 1, 2, 3, \quad (6)$$

with cross-covariance

$$c_{km}(\mathbf{h}) = \text{cov}\{z_k(\mathbf{s}), z_m(\mathbf{s} + \mathbf{h})\} = \rho_k \rho_m \int_{\mathbb{R}^2} K_k(\mathbf{u})K_m(\mathbf{u} - \mathbf{h})\mathbf{d}\mathbf{u}, \quad k \neq m, \quad (7)$$

admitting positive, negative, and zero cross-correlation, depending on the value of $\rho_{km} = \rho_k \rho_m$. We note that both $\boldsymbol{\rho}$ and $-\boldsymbol{\rho}$ will yield identical forms in (7); however, this nonidentifiability is easily remedied by constraining $\rho_1 > 0$ with no loss in modeling flexibility.

A wide variety of spatial covariance structures can be obtained in (6) and (7) through an appropriate choice of smoothing kernels in (5). The Gaussian kernel is commonly employed (Higdon, 1998; Higdon, 2002; Calder, 2007) and corresponds to a Gaussian correlation function. We allow for anisotropy and added flexibility through two generalizations, the first based on the density of a bivariate t-distribution

$$K_l(\mathbf{u}) \propto \sigma_l \left(1 + \frac{u_x^2}{\nu_l \phi_{lx}} + \frac{u_y^2}{\nu_l \phi_{ly}}\right)^{-(\nu_l+2)/2}, \quad (8)$$

and the second based on the Matérn covariance function

$$K_l(\mathbf{u}) \propto \sigma_l \left(\frac{\nu_l u_x^2}{\phi_{lx}} + \frac{\nu_l u_y^2}{\phi_{ly}} \right)^{\nu_l/2} k_{\nu_l} \left(2 \sqrt{\frac{\nu_l u_x^2}{\phi_{lx}} + \frac{\nu_l u_y^2}{\phi_{ly}}} \right), \quad (9)$$

where $k_{\nu_l}(\cdot)$ is a modified Bessel function of the second kind having order ν_l ; $\sigma_l > 0$ is a scale parameter; $\phi_{lx} > 0$ and $\phi_{ly} > 0$ are componentwise spatial range parameters; and $\nu_l > 0$ controls the length of the tails in (8) and the smoothness of process realizations associated with (9). A Gaussian kernel arises as a limiting form in both (8) and (9) as $\nu_l \rightarrow \infty$.

2.3.2 Dimension Reduction

To facilitate computation with data observed at a large number of locations, we work with (5) through a dimension-reduced discrete sum approximation. We let $\{\mathbf{A}_1, \dots, \mathbf{A}_J\}$ denote a collection of J disjoint and bounded sets covering the spatial domain $\mathbb{D}$, with corresponding centroids $\{\mathbf{t}_1, \dots, \mathbf{t}_J\}$, and define $A^* = \bigcup_{j=1}^J A_j$ so that $\mathbb{D} \subset A^* \subset \mathbb{R}^2$. For $\mathbf{s} \in \mathbb{D}$, we first approximate (5) by restricting integration to the bounded set A^*

$$\mathbf{z}(\mathbf{s}) \approx \boldsymbol{\mu}_z + \int_{A^*} \mathbf{K}(\mathbf{u} - \mathbf{s}) \mathbf{T} \mathbf{w}(\mathbf{u}) d\mathbf{u} = \sum_{j=1}^J \int_{A_j} \mathbf{K}(\mathbf{u} - \mathbf{s}) \mathbf{T} \mathbf{w}(\mathbf{u}) d\mathbf{u}. \quad (10)$$

Assuming further that $\mathbf{K}(\mathbf{u} - \mathbf{s}) \approx \mathbf{K}(\mathbf{t}_j - \mathbf{s})$ for all $\mathbf{u} \in \mathbf{A}_j$ $j = 1, \dots, J$, yields the discrete process convolution

$$\mathbf{z}(\mathbf{s}) \approx \mathbf{z}^*(\mathbf{s}) = \boldsymbol{\mu}_z + \sum_{j=1}^J \mathbf{K}(\mathbf{t}_j - \mathbf{s}) \mathbf{X}_j, \quad (11)$$

where $\mathbf{X}_j = \mathbf{T} \int_{A_j} \mathbf{w}(\mathbf{u}) d\mathbf{u}$ so that

$$\mathbf{X}_j \stackrel{ind}{\sim} MVN_3(\mathbf{0}, |A_j| \mathbf{T} \mathbf{T}'), \quad j = 1, \dots, J, \quad (12)$$

where $|A_j|$ denotes the area of $\mathbf{A}_j$. For application to our analysis in Section 4, we will assume that $\{\mathbf{A}_1, \dots, \mathbf{A}_J\}$ is a collection of square subregions, each having area ϵ^2 , so that

$\mathbf{t}_1, \dots, \mathbf{t}_J$ represent locations on a regular grid with spacing $\epsilon > 0$. Extensions to adaptive methods, allowing for a data driven choice of the support points $\mathbf{t}_1, \dots, \mathbf{t}_J$, are discussed in Section 5. In practice, the value of J , and hence the number of latent variables characterizing the dimension-reduced process (11), is determined by computational constraints.

3. Bayesian Inference and Computation

3.1 Likelihood and Prior Distribution

Letting $\mathbf{N}$ and $\mathbf{Y}$ denote the observed interval counts and log-height measurements respectively, the likelihood function associated with our model formulation takes the form

$$L(\boldsymbol{\theta}|\mathbf{N}, \mathbf{Y}) = \prod_{i=1}^n \left[\prod_{j=1}^{M_1} P_N(N_{ij}|\mu_{N_{ij}}) \right] \times \left[\prod_{j=1}^{M_2} P_Y(Y_{ij}|\mu_{Y_{ij}}, \sigma_\epsilon^2) \right]$$

where $P_Y(\cdot|\cdot, \cdot)$ and $P_N(\cdot|\cdot)$ denote the density and probability mass function of the Gaussian and Poisson distributions respectively;

$$\mu_{N_{ij}} = \int_{t_{j-1}^{(N)}}^{t_j^{(N)}} \lambda_i\{u|x_{\mathbf{s}_i}^{(H)}(u)\}du, \quad (13)$$

and $\mu_{Y_{ij}} = x_{\mathbf{s}_i}(t_j^{(H)})$ is defined by (3).

Our Bayesian model specifications are made complete upon assigning a prior distribution $\pi(\boldsymbol{\theta})$ which, based on the dimension-reduced process (11), factorizes as

$$\begin{aligned} \pi(\boldsymbol{\theta}) &= \pi(\beta)\pi(\sigma_\epsilon^2) \left[\prod_{i=1}^n \pi(\omega_i|\eta) \right] \pi(\eta) \left[\prod_{j=1}^{M_1} \pi(\lambda_j) \right] \\ &\times \left[\prod_{l=1}^J \pi(\mathbf{X}_l|\boldsymbol{\rho}) \right] \pi(\boldsymbol{\rho}) \left[\prod_{l=1}^3 \pi(\sigma_l)\pi(\phi_{lx})\pi(\phi_{ly})\pi(\nu_l)\pi(\mu_{Z_l}) \right], \end{aligned}$$

where $\pi(\omega_i|\eta) \propto \omega_i^{\eta-1} \exp(-\eta\omega_i)$ is the density of the gamma distributed frailties, and $\pi(\mathbf{X}_1|\boldsymbol{\rho})$ is a trivariate Gaussian density associated with the increments (12). Specific forms for the remaining priors are discussed in Section 4 where the application is considered. The conditional independence graph illustrating the relationships between model components is depicted in Figure 2. Inference is based on Monte Carlo samples drawn from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{N}, \mathbf{Y}) \propto L(\boldsymbol{\theta}|\mathbf{N}, \mathbf{Y})\pi(\boldsymbol{\theta})$, which we obtain using an MCMC algorithm described in the Section 3.2.

For comparison of spatial models based on (8) and (9), we adopt an approach based on the conditional predictive ordinate (CPO) (see e.g. Gelfand and Dey 1994). In our setting, we define these as $CPO_{Y_{ij}} = f(Y_{ij}|\mathbf{Y}_{-ij}, \mathbf{N})$, $i = 1, \dots, n$, $j = 1, \dots, M_1$, and $CPO_{N_{ij}} = f(N_{ij}|\mathbf{N}_{-ij}, \mathbf{Y})$, $i = 1, \dots, n$, $j = 1, \dots, M_2$, where $\mathbf{Y}_{-ij}$ ($\mathbf{N}_{-ij}$) represents the height (infection) data without the observation Y_{ij} (N_{ij}), and $f(\cdot|\cdot)$ denotes the cross validation predictive density. For a given observation, higher values of the corresponding CPO indicate greater support for the model at a local (observational) level. An overall summary for global model comparison is based on the pseudo marginal loglikelihood (PSML) defined as

$$PSML = \sum_{i=1}^n \left[\sum_{j=1}^{M_1} \log(CPO_{Y_{ij}}) + \sum_{j=1}^{M_2} \log(CPO_{N_{ij}}) \right], \quad (14)$$

with higher values indicating preferred models. We estimate each $CPO_{Y_{ij}}$ and $CPO_{N_{ij}}$ using the harmonic mean estimator proposed by Gelfand and Dey (1994), which is readily computed using samples from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{N}, \mathbf{Y})$ (see Web Appendix).

3.2 Computational Implementation based on Hybrid Monte Carlo

In what follows we highlight the most important aspects of the algorithm used for sampling the posterior distribution. A detailed specification of the algorithm, including discussion of reparameterizations that accelerate MCMC convergence, is outlined in the Web Appendix. In implementing a componentwise Metropolis-Hastings sampler, we have found

that Monte Carlo updating of the latent variables $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_J)$, defined in (12), requires close consideration. In particular, random walk updating of these parameters leads to slow movement through the target distribution, despite careful tuning of proposal distributions, reparameterizations, and implementations based on various blocking schemes. As a remedy, we employ a hybrid Monte Carlo algorithm (Gustafson 1997; Ishwaran 1999), which is designed to suppress inefficient random walk behavior and promote rapid mixing of the Markov chain.

The hybrid algorithm is based on the combination of a stochastic step with a preset number of deterministic steps that represent a discretization of Hamiltonian dynamics. Each iteration of our algorithm involves a block update of $\mathbf{X}$ based on the hybrid procedure. Letting $\pi_{\mathbf{X}}(\cdot)$ denote the density of the target distribution $[\mathbf{X}|\mathbf{N}, \mathbf{Y}, \boldsymbol{\theta}_{-\mathbf{X}}]$, and $\mathbf{X}^*$ the current value in the Markov chain for $\mathbf{X}$, the hybrid update, based on a step size $\delta > 0$, proceeds as follows:

1. *Simulate auxiliary variables $\mathbf{U}^* \sim MVN_{3J}(\mathbf{0}, \mathbf{I})$.*

Let $\mathbf{X}^{(0)} = \mathbf{X}^$ and $\mathbf{U}^{(0)} = \mathbf{U}^* + \frac{\delta}{2} \nabla \log \pi_{\mathbf{X}}(\mathbf{X}^*)$*

2. *For $l = 1, \dots, L$, let*

$$\mathbf{X}^{(l)} = \mathbf{X}^{(l-1)} + \delta \mathbf{U}^{(l-1)}$$

$$\mathbf{U}^{(l)} = \mathbf{U}^{(l-1)} + \delta_l \nabla \log \pi_{\mathbf{X}}(\mathbf{X}^{(l)})$$

where $\delta_l = \delta$ for $l < L$ and $\delta_L = \frac{\delta}{2}$.

3. *Accept $\mathbf{X}^{(L)}$ as the new state for $\mathbf{X}$ with probability*

$$p = \min \left(\frac{\pi_{\mathbf{X}}(\mathbf{X}^{(L)})}{\pi_{\mathbf{X}}(\mathbf{X}^*)} \exp \left\{ -\frac{1}{2} \left(\mathbf{U}^{(L)'} \mathbf{U}^{(L)} - \mathbf{U}^{*'} \mathbf{U}^* \right) \right\}, 1 \right)$$

else remain in the current state $\mathbf{X}^$ with probability $1 - p$.*

Each hybrid update requires $L + 1$ evaluations of the gradient vector $\nabla \log \pi_{\mathbf{X}}(\cdot)$ (see Web Appendix for the derivation) and we have found that setting $L = 50$ works well in the current setting.

Upon updating $\mathbf{X}$, the sampler visits the remaining components of $\boldsymbol{\theta}$ using a sequence of random walk Metropolis-Hastings steps and Gibbs updates. While it is possible to employ the hybrid algorithm for updating these remaining components, such updates are not required for successful posterior sampling. Candidate distributions for random walk updates are tuned to yield acceptance rates of between 20 and 50 percent; whereas, δ is chosen so that the hybrid step has an acceptance rate of between 75 and 90 percent.

4. Study of Pine Weevil Infestation

In this section we apply our methods to the weevil infestation data described in Section 1. The data were collected by the British Columbia Ministry of Forests, over an eleven year period, beginning in 1984. The study region, depicted in Figure 1, was planted in the Spring of 1984 using 1-year-old container seedlings. Each tree was inspected for infection on an annual basis, and the number of weevil attacks was recorded. The infection data thus arise in a panel structure of $M_1 = 11$ followup intervals, generating annual infection counts. Total tree height (in centimeters) was measured at $M_2 = 4$ time points (years after planting) with $t_1^{(H)} = 0$, $t_2^{(H)} = 3$, $t_3^{(H)} = 6$ and $t_4^{(H)} = 10$. We thus fit our models to a total of $(M_1 + M_2) * n$ observations, with $n = 4330$ trees resulting in a total of 64,950 observations, measured over space and time.

Two models corresponding to the smoothing kernels (8) and (9) are fit. In implementing the spatial process (11), we take $J = 63$ and $\epsilon = 15$ meters, with the resulting grid of support points depicted in Figure 1. Priors for the spatial range parameters are based on this grid spacing, where we set $\phi_{lx} = \epsilon^2 \psi_{lx}$ and $\phi_{ly} = \epsilon^2 \psi_{ly}$ with $\psi_{lx}, \psi_{ly} \sim \text{Unif}(0.05, 1)$.

Other priors for parameters of the spatial process model are specified as $\nu_l \stackrel{iid}{\sim} \text{Unif}(0.05, 10)$, $\mu_{z_l} \stackrel{iid}{\sim} N(0, 100)$, $\rho_1 \sim \text{Unif}(0.0, .99)$, $\rho_2, \rho_3 \sim \text{Unif}(-0.99, 0.99)$, and the variance components are assigned unit mean inverse-gamma priors $\sigma_l^2 \stackrel{iid}{\sim} IG(2, 1)$. The remaining model parameters are assigned fairly vague priors that are taken to be conditionally conjugate whenever possible: $\beta \sim N(0, 100)$, $\eta \sim \text{Gamma}(.01, .01)$, $\lambda_j \stackrel{iid}{\sim} \text{Gamma}(.01, .01)$, $\sigma_\epsilon^2 \sim IG(2, 1)$. For each model, we ran four parallel chains of the MCMC algorithm described in Section 3.2 for a total of 25,000 iterations. Upon examining the resulting output for stationarity, a burn-in of 3,000 iterations was chosen with the remaining samples used for posterior inference.

To compare the two models based on either (8) or (9) we compute the PSML criteria (14) and obtain values of $PSML = -12045$ and $PSML = -12048$ respectively. The small difference indicates roughly equal performance of both models and; moreover, illustrates the general difficulty of discriminating between competing spatial covariance structures (Ecker and Gelfand, 1997), particularly when these are used to model a latent process. As both model fits yield very similar results, we present posterior summaries for the model based on the Matérn spatial structure in more detail. Figure 3, panel (a) displays the posterior mean and the pointwise 95% highest posterior density (HPD) interval for the baseline hazard of infection $\lambda_0(t)$. Attacks on newly planted seedlings are rare, with the estimated hazard being extremely low during the first four years after planting. This is followed by a monotone increasing hazard to year eight, and then by a leveling off in the risk as the plantation reaches a steady state.

Table 2 presents posterior summaries for the remaining model parameters. The posterior distribution for β reveals a positive association between the hazard of infection and height, with taller trees having a higher rate of infection. Specifically, the posterior mean (95% HPD) relative risk of infection associated with a unit increase in height (cm on the log scale) is 1.73 (1.07, 2.48). This result is consistent with the notion that weevils prefer those trees having the longest and thickest leaders (King et al. 1997). While height is indeed

important in explaining the variability in infection, the posterior credible interval for the frailty variance η^{-1} indicates a non-negligible degree of residual heterogeneity. It is likely that this can be attributed, in large degree, to genetic differences among trees. The inclusion of genetic information and the resulting decrease in residual heterogeneity is an interesting avenue for further exploration.

Moving next to the parameters governing spatial variability and height growth, we examine posterior summaries for the six spatial range parameters (ψ_{lx}, ψ_{ly}) , $l = 1, 2, 3$. Comparing estimates within each pair reveals roughly equal values for scaling in each direction, and hence isotropy may be a reasonable assumption for spatial modeling of each growth parameter. The data are not particularly informative regarding the smoothness parameters ν_l , with relatively wide credible intervals and little prior-to-posterior movement in each case. Examining posterior summaries for the cross-covariance parameters ρ_{12} , ρ_{13} , and ρ_{23} , it seems that a multivariate spatial structure, as opposed to the rather strong assumption of independent spatial processes is indeed warranted. In particular, the posterior distribution for ρ_{12} indicates considerable negative dependence between $b_1(\mathbf{s})$ and $b_2(\mathbf{s})$. This negative relationship is further illustrated in Figure 4, which displays interpolated surfaces for the asymptotic height process $b_1(\mathbf{s})$ as well as $b_2(\mathbf{s})$ (see Web Figure 1 for $b_3(\mathbf{s})$). Regions of both high and low growth are visible. The apparent hot spots and low spots for asymptotic height are particularly informative, and can be used to guide further study with the goal of identifying unknown factors related to height growth. Two specific locations are marked in Figure 4 with the symbols ‘L’ and ‘H’, corresponding to regions of low and high growth respectively. Figure 3, panel (b) displays the growth trajectories for the trees located at these two positions. A substantial difference in growth rates is apparent and; moreover, this difference helps to illustrate the range of variability in growth patterns observed across the plantation.

4.1 Goodness-of-fit to the Pine Weevil Data

To assess the fit of our model we work with the posterior predictive distribution $[\mathbf{N}^{(rep)}, \mathbf{Y}^{(rep)} | \mathbf{N}, \mathbf{Y}]$. We compare replicate data, simulated from this distribution, to the observed data through specifically chosen discrepancy functions $T(\mathbf{N}, \mathbf{Y}, \boldsymbol{\theta})$ (Gelman et al. 1996). First, we use a discrepancy based on the deviance $T(\mathbf{N}, \mathbf{Y}, \boldsymbol{\theta}) = -2 \log L(\boldsymbol{\theta} | \mathbf{N}, \mathbf{Y})$ as an *overall* check on model adequacy. In this case, the posterior predictive p-value for assessing model adequacy is given by $p_b = Pr(T(\mathbf{N}^{(rep)}, \mathbf{Y}^{(rep)}, \boldsymbol{\theta}) \geq T(\mathbf{N}, \mathbf{Y}, \boldsymbol{\theta}) | \mathbf{N}, \mathbf{Y})$, where the probability is taken with respect to the joint posterior $[\mathbf{N}^{(rep)}, \mathbf{Y}^{(rep)}, \boldsymbol{\theta} | \mathbf{N}, \mathbf{Y}]$, with values close to zero or one indicating a lack of fit. Having simulating values $\mathbf{N}^{(rep,k)}, \mathbf{Y}^{(rep,k)}, \boldsymbol{\theta}^{(k)}, k = 1, \dots, 5000$ from the joint posterior, Figure 5 panel (a) displays a scatterplot of the pairs $\{(T(\mathbf{N}, \mathbf{Y}, \boldsymbol{\theta}^{(k)}), T(\mathbf{N}^{(rep,k)}, \mathbf{Y}^{(rep,k)}, \boldsymbol{\theta}^{(k)}))\}, k = 1 \dots, 5000\}$. The posterior predictive p-value is computed as the proportion of these pairs for which $T(\mathbf{N}^{(rep,k)}, \mathbf{Y}^{(rep,k)}, \boldsymbol{\theta}^{(k)})$ exceeds $T(\mathbf{N}, \mathbf{Y}, \boldsymbol{\theta}^{(k)})$, and is found to be $p_b = 0.650$, so that a lack of overall fit is not indicated. In Figure 5 panel (b) we plot the cumulative infection counts, summed over all trees, in each year and compare with the corresponding 95% predictive intervals. Again, no lack of fit is evident with the predictive replicates capturing the observed changes in cumulative infection effectively. Finally, we check the adequacy of our spatial modeling of the height data through the spatial correlogram based on Moran's I (Cliff and Ord, 1973). For the spatial height measurements obtained in year $t_j^{(H)}$, this function is defined over a set of increasing distance classes or spatial lags $d_1, \dots, d_K$ by

$$I_j(d_p) = \frac{n \sum_{\{i,k \in \mathbf{C}_p\}} (Y_{ij} - \bar{Y}_j)(Y_{kj} - \bar{Y}_j)}{|\mathbf{C}_p| \sum_{i=1}^n (Y_{ij} - \bar{Y}_j)^2}, \quad p = 1, \dots, K; \quad j = 1, \dots, M_2$$

where $\mathbf{C}_p = \{(i, k) : \|\mathbf{s}_i - \mathbf{s}_k\| \in d_p\}$ a set having $|\mathbf{C}_p|$ elements, and $\bar{Y}_j = \frac{1}{n} \sum_{i=1}^n Y_{ij}$. In the absence of spatial correlation, $I_j(\cdot)$ has an expected value close to zero; whereas, positive and negative values indicate positive and negative spatial correlation respectively. For each of the

$M_2 = 4$ years where heights were measured, the observed values of the spatial correlogram are compared with the 95% posterior predictive intervals in Figure 5, panels (c) through (f). The correlograms are based on eight distance classes each having a width of 10 meters. The model captures the spatial dependence reasonably well. In particular, comparing the correlograms across the four years of measurement, we note that the spatial correlation structure in the observed data exhibits a changing pattern over time. The predictive fits clearly indicate the ability of our model to track these changes.

5. Discussion

We have proposed a novel joint spatial model for recurrent infection and growth, applicable for analysis of longitudinal studies in forestry where the hazard of infection depends on tree growth over time. We develop a spatially dynamic nonlinear model for growth trajectories, and incorporate the trajectory as a spatially-varying covariate in a time-dependent Cox model relating infection and growth. Inference was conducted in a Bayesian framework with implementation based on a hybrid Markov chain Monte Carlo algorithm.

There are several directions for future work that are currently being investigated. First, alternative approaches for representing the hazard function would be useful to explore. The time-dependent Cox model (4) is a standard representation; however, other forms, for example, a form based on the time-dependent accelerated failure time model can also be employed. Following Tseng et al. (2005), a specification based on an accelerated failure time model takes the form

$$\lambda_i\{t|x_{\mathbf{s}_i}^{(H)}(t)\} = \lambda_0 \left[\omega_i \int_0^t \exp\{\beta x_{\mathbf{s}_i}(u)\} du \right] \omega_i \exp\{\beta x_{\mathbf{s}_i}(t)\}$$

and yields an interesting alternative model for comparison.

We are also investigating *adaptive* discrete process convolutions, where the grid locations $\mathbf{t}_1, \dots, \mathbf{t}_J$ supporting the discrete process are estimated from the data rather than fixed a priori. This can be achieved by adding another level to the hierarchical formulation and modeling the grid itself as a realization from a spatial point process. The adaptive model may result in greater capability to capture fine scale spatial structure and this may be useful in other applications. Nevertheless, issues regarding model identifiability and efficient computation with these more flexible models need careful consideration.

ACKNOWLEDGEMENTS

Research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The author thanks Dr. John King of the British Columbia Ministry of Forests for providing the data.

List of Tables

1. Posterior mean and 95% highest posterior density interval for select model parameters.

List of Figures

1. Locations of trees within the plantation with grid points (larger circles) for discrete kernel convolution superimposed.
2. Directed acyclic graph illustrating model structure and conditional independence assumptions. Solid boxes represent observed quantities, circles represent stochastic unknowns and dashed boxes represent repeated structures.
3. Panel (a) depicts the posterior mean (solid curve) and 95% highest posterior density (HPD) interval (dashed curve) for the baseline hazard of infection; (b) posterior mean and 95% HPD intervals for growth trajectories of the two specific trees whose positions are marked in Figure 4 with an 'H' (grey curves) and an 'L' (black curves).
4. Image plot of the posterior mean interpolated surfaces for $b_1(\mathbf{s})$ (a) and $b_2(\mathbf{s})$ (b). Individual growth trajectories for the two trees marked 'H' and 'L' are depicted in Figure 3, panel (b).
5. Posterior predictive model checks: (a) predictive versus realized values of the deviance; (b) cumulative infection totals at each year, where circles represent observed values and dashed lines represent 95% posterior predictive intervals. Panels (c) through (f) give the spatial correlograms for (log) height at each of the four years where heights were measured.

Table 1*Posterior mean and 95% highest posterior density interval for select model parameters*

Parameter	Posterior Mean	95% HPD
β	0.524	(0.098, 0.924)
η^{-1}	0.722	(0.613, 0.834)
σ_{ϵ}^2	0.081	(0.079, 0.083)
ψ_{1x}	0.747	(0.564, 0.924)
ψ_{1y}	0.678	(0.536, 0.850)
ψ_{2x}	0.785	(0.540, 1.000)
ψ_{2y}	0.865	(0.657, 1.000)
ψ_{3x}	0.240	(0.050, 0.606)
ψ_{3y}	0.256	(0.051, 0.631)
ν_1	7.342	(3.786, 10.000)
ν_2	6.113	(1.766, 9.996)
ν_3	3.008	(0.135, 9.096)
ρ_{12}	-0.969	(-0.980, -0.941)
ρ_{13}	-0.454	(-0.851, 0.020)
ρ_{23}	0.454	(-0.020, 0.844)
σ_1^2	0.190	(0.109, 0.285)
σ_2^2	0.193	(0.095, 0.308)
σ_3^2	0.484	(0.149, 0.958)
μ_{z_1}	1.655	(1.653, 1.657)
μ_{z_2}	-0.934	(-0.941, -0.926)
μ_{z_3}	-1.257	(-1.276, -1.238)

Figure 1. Locations of trees within the plantation with grid points (larger circles) for discrete kernel convolution superimposed.

Figure 2. Directed acyclic graph illustrating model structure and conditional independence assumptions. Solid boxes represent observed quantities, circles represent stochastic unknowns and dashed boxes represent repeated structures.

Figure 3. (a) Posterior mean (solid curve) and 95% highest posterior density (HPD) interval (dashed curve) for the baseline hazard of infection; (b) posterior mean and 95% HPD intervals for growth trajectories of the two specific trees whose positions are marked in Figure 4 with an 'H' (grey curves) and an 'L' (black curves).

Figure 4. Image plot of the posterior mean interpolated surfaces for $b_1(\mathbf{s})$ (a) and $b_2(\mathbf{s})$ (b). Individual growth trajectories for the two trees marked 'H' and 'L' are depicted in Figure 3, panel (b).

Figure 5. Posterior predictive model checks: (a) predictive versus realized values of the deviance; (b) cumulative infection totals at each year, where circles represent observed values and dashed lines represent 95% posterior predictive intervals. Panels (c) through (f) give the spatial correlograms for (log) height at each of the four years where heights were measured.