A two-stage model for incidence and prevalence in point-level spatial count data

Virginia Recta
Department of Statistics
Pennsylvania State University
vfr100@psu.edu

Murali Haran
Department of Statistics
Pennsylvania State University
mharan@stat.psu.edu

James L. Rosenberger
Department of Statistics
Pennsylvania State University
jlr@stat.psu.edu

Abstract

We consider the problem of modeling point-level spatial count data with a large number of zeros. We develop a model that is compatible with scientific assumptions about the underlying data generating process. We utilize a two-stage spatial generalized linear mixed model framework for the counts, modeling incidence, resulting in 0-1 outcomes, and prevalence, resulting in positive counts, as separate but dependent processes, and utilize a Gaussian process model for characterizing the underlying spatial dependence. We describe a Bayesian approach, and study

several variants of our two-stage model. We fit the models via Markov chain Monte Carlo (MCMC) methods. We study several MCMC algorithms, including a version of the Langevin-Hastings algorithm, for exploring the complicated posterior distribution efficiently, and recommend an algorithm that is fairly efficient. Finally, we demonstrate the application of our modeling and computational approach on both simulated data and real data from an ecological field survey.

1 Introduction

Spatial count data arise frequently in a number of disciplines. There is an increasing awareness of the need to use models that account for the spatial dependence commonly inherent in such data. For example, in the study of insect populations, geostatistical tools have been used to capture the degree of spatial dependence that is present in most populations (Legendre and Fortin, 1989; Rossi et al., 1992; Schotzko and O'Keeffe, 1989; Schotzko and Smith, 1991; Williams et al., 1992). Advances in Global Positioning Systems (GPS) technology have permitted rapid and accurate capture of field data at finer scales of resolution and greater sampling intensity. For instance, Blom and Fleischer (2001) describe a study of the spatial dynamics of Colorado potato beetle (CPB) populations in potato fields at the level of density per meter. One complication encountered in their study, however, was that a substantial proportion of the observations were zeros. The spatial distribution and histogram of the raw observations from the CPB study are displayed in Figures 1 and 2. From a scientific point of view, the distribution may be seen as a manifestation of two biological processes: incidence, which is a binary (absence/presence) variable, and prevalence, which is a count variable. Note that in this framework, the count variable is only observed when the binary variable indicates presence.

Studying the incidence and prevalence processes separately but simultaneously

allows for a model that is compatible with the hypothesized underlying data generating process. Such a model can also be useful in characterizing relationships between each of the processes and potential predictors. At varying times and insect stages, specific interest may also be on various functionals of the distribution, in addition to the usual spatial predictions; sample-based inference is easily extended to handle such problems as well.

Agarwal et al. (2002) describe a mixture model for zero-inflated areal (spatially aggregated) data where absence (zeros) appear in a particular region with probability p and, with probability $1-p$, a Poisson random variate is generated for that region. Presence-absence is modeled via a logistic regression, while the Poisson mean is modeled via a standard log-linear model. Spatial dependence for the areal data is modeled via a Gaussian Markov random field model imposed on the random effect terms in the log-linear model. This is very well suited to areal data, though it is perhaps worth noting that the interpretation of parameters is specific to the configuration of sites (or subregions) on which the model is defined. Rathbun and Fei (2006) propose a probit model for zero-inflated spatial data on a continuous spatial domain, developed along the lines of the spatial probit model described in De Oliveira (2000), with excess zeros generated at a given site if the realization of a Gaussian process falls below a threshold. This model is well suited to problems where the zero inflation arises largely as a result of detection limits. In our motivating example, however, the zeros are true zeros, that is, zeros in the data truly indicate absence.

— INSERT FIGURES 1,2 ABOUT HERE —

In contrast, two-stage or two-part models are appropriate for studying the process in stages: first, we study the process that produces zero versus non-zero outcomes, and second, we can model the count process conditional on positive outcomes. In the longitudinal setting, Olsen and Schafer (2001) cite several examples for which two-stage models are ideally suited: adolescent substance abuse, dividend income, and

expenditures on durable goods and medical care. Ver Hoef and Jansen (2007) describe a two-stage 'hurdle' model in the context of modeling areal data. In two-stage models, zeros are real outcomes, and do not simply represent insufficient information. Hence, the two-stage conditional specification is well suited to the type of spatial phenomena that we are considering in this research. Here we describe a model that utilizes the spatial generalized linear mixed model framework described in Diggle et al. (1998) (also see Diggle and Ribeiro (2007); Haran (2011)). We break the process into two parts, one for incidence and one for prevalence, with each part modeled via a generalized linear model, with spatial dependence specified via a Gaussian process model for the random effects associated with each of the two processes. Our model formulation allows us to explore various specifications, including a model that allows for a cross-covariance between the two processes. Our inference is based on the resulting posterior distribution, which can be estimated using Markov chain Monte Carlo (MCMC). Since constructing efficient MCMC algorithms is challenging due to the complicated posterior distribution and dependence among the parameters, we study several MCMC algorithms, and recommend a fairly efficient version of the Langevin-Hastings MCMC algorithm that appears to be reasonably robust to differences among data sets.

The rest of our paper is organized as follows. In Section 2 we describe our two-stage spatial model, and in Section 3 we discuss ways to resolve the challenging computational problems. In Section 4, we compare and contrast variants of our two-stage model, and apply our approach to both simulated and real data sets. Finally, we conclude with a discussion of our results in Section 5.

2 A two-stage spatial model

We begin with some background — a brief description of spatial generalized linear mixed models, followed by a description of two-stage modeling in a non-spatial framework. We then describe our two-stage spatial models.

The models we develop here differ from the zero-inflated Poisson (ZIP) models in Agarwal et al. (2002) and Rathbun and Fei (2006) with respect to model construction as well as incorporation of spatial dependence. In the ZIP models, the zero observations may arise from both the binary (incidence) and count (prevalence) processes. In contrast, the two-stage model presented here completely specifies and separates the binary from the count processes. Also, while their framework allows in principle for modeling spatial dependence in both the count and binary processes, Agarwal et al. (2002) explore spatial dependence in the log-normal (count) part but not in the incidence (binary) part. Rathbun and Fei (2006) use a probit model with spatial random effects for the incidence part, where the species of interest is observed only if the combined effects of environmental conditions overcome a threshold; the prevalence process is modeled without spatial random effects. Gschlößl and Czado (2008) consider spatial ZIP models for areal data using a proper Gaussian conditional autoregression, and Fernandes et al. (2009) propose a model for zero-inflated spatio-temporal processes, including the case of continuous observations with spatial random effects following a zero-mean Gaussian process.

Ver Hoef and Jansen (2007) adopt an approach similar to ours in a 'hurdle model' used to investigate haul-out patterns of harbor seals on glacial ice, and describe a very nice comparative study of the hurdle model to other spatial ZIP models. In contrast to Ver Hoef and Jansen (2007), who consider hurdle models in the context of Gaussian Markov random field models for areal (aggregated) or lattice space-time data, in our model we assume that the zero-inflated observations are geostatistical.

In particular, we assume they arise from a bivariate stochastic process on a continuous spatial domain. This model is useful in many ecological and biological settings where such data are common. Also, our model allows us to interpolate realizations in places where there are no observations, while also giving us the ability to study the dependence in the spatial process since covariance function parameters have a more natural interpretation and do not rely on definitions of sub-regions and neighborhoods, which can be arbitrary. We also include a mechanism to relate the two parts of the model via a cross-covariance between the spatial random effects. Computation for these models can be challenging; we therefore describe some approaches in Section 3 to overcome these challenges.

2.1 Spatial generalized linear mixed models

Diggle et al. (1998) propose an approach for modeling data where the known sampling mechanism for the data is non-Gaussian. These spatial generalized linear mixed models (SGLMMs) utilize the generalized linear model framework (McCullagh and Nelder, 1983) for spatially associated data. The spatial dependence (the error structure) for SGLMMs can be modeled via Gaussian processes for point-level (geostatistical) data as described in the seminal paper by Diggle et al. (1998), and a similar framework can also be used for areal or lattice data with a Gaussian Markov random field (GMRF) model (cf. Banerjee et al., 2004; Haran, 2011; Rue and Held, 2005). Following the notation in Diggle et al. (1998), the model can be described as follows:

- Let $\{S(x) : x \in D\}$, where $D \subset \mathbb{R}^d$, be a zero-mean Gaussian process with the covariance of the spatial process specified via a valid (positive definite) covariance function, say from the Matérn family (cf. Handcock and Stein, 1993).
- Conditionally on $S(x)$, the random variables $\{Y(x) : x \in D\}$ are mutually

independent, with distributions $f(y \mid S(x)) = f(y \mid M(x))$ where $M(x) = E(Y(x) \mid S(x))$.

- Define $h(M(x)) = \mathbf{d}(x)\boldsymbol{\beta} + S(x)$ for some known link function h (say log or logit), vector of explanatory variables $\mathbf{d}(x)$ and parameters $\boldsymbol{\beta}$.

The $\{Y(x) : x \in D\}$ process can then be modeled, conditional on the spatial random effects S, as Poisson or Negative Binomial random variables when modeling counts or as Bernoulli random variables when modeling binary outcomes, for instance. Such models are convenient for predicting non-linear functionals of realized values such as the maximum value over a region, or the probability of exceeding a specified threshold under possibly non-Gaussian realizations. This structure allows for a very rich class of models that can be used for a variety of spatial processes. Once priors are specified for the parameters of the covariance function and $\boldsymbol{\beta}$, MCMC techniques can be used to estimate and make inferences about the parameters, predict realizations at arbitrary locations, and estimate non-linear functionals of the posterior distribution.

2.2 Two-stage models

Consider the general problem of modeling variables with two components. Olsen and Schafer (2001), for instance, describe a model for semicontinuous variables — variables that have a portion of responses equal to a single value (usually zero) and a continuous, often skewed, distribution for the remaining values. Olsen and Schafer (2001) consider such variables in a longitudinal setting. In modeling semicontinuous longitudinal data, the semicontinuous response, Y_{ij}, can be coded into two variables,

$$U_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \neq 0 \\ 0 & \text{if } Y_{ij} = 0 \end{cases} \tag{1}$$

$$V_{ij} = \begin{cases} g(Y_{ij}) & \text{if } Y_{ij} \neq 0 \\ \text{irrelevant} & \text{if } Y_{ij} = 0 \end{cases} \tag{2}$$

where $j = 1, \ldots, n_i$ indexes the time points for individual i where $i = 1, \ldots, m$, and g is a monotone increasing function (say log). They then fit a two-stage random-effects model, one for the logit probability $U_{ij} = 1$ and one for the mean conditional response $E(V_{ij}|U_{ij} = 1)$. This approach allows for a different set of covariates for each part of the model, that is, a set of covariates for the probability of nonzero response and another set for the mean of nonzero responses. The same modeling approach applies naturally to model the process that produces the zero-inflated count data described in Section 1. At the same time, a joint distribution for the random coefficients from each part provides a mechanism for relating the two parts of the model. This enables us to specify one viable model for two separate but related phenomena: the binary indicator of whether there is at least one occurrence, and the distribution of positive occurrences.

2.3 Spatial two-stage models

We now describe a two-stage model for spatial data on a continuous spatial domain. We begin by describing the model in some generality, then provide a more specific version of it for the purposes of our study. Consider the response at location x_i, $Y_i = Y(x_i), i = 1, \ldots, n$. We decompose Y_i into two variables, a binary part and a discrete (or continuous) part in similar fashion to (1) and (2), with $U_i = 1$ if $Y_i > 0$, and $U_i = 0$ if $Y_i = 0$. Also, $V_i = Y_i$ if $Y_i > 0$, with V_i irrelevant whenever $Y_i = 0$. There are n observations for U, of which $n_1 \leq n$ are equal to 1, and the rest are 0. For convenience, we order the data so that the 1's are the first n_1 observations. There are n_1 observations for V, corresponding to the first n_1 observations of U. To incorporate spatial dependence, we condition the U and V processes on the

Gaussian processes S and Z respectively. Let $\mathbf{0}_n, \mathbf{0}_{n_1}$ be vectors of n and n_1 zeros respectively, and $S(x_i), Z(x_i)$ denote the processes at the location x_i. Furthermore, let $\mathbf{S} = (S(x_1), \ldots, S(x_n))^T$ and $\mathbf{Z} = (Z(x_1), \ldots, Z(x_{n_1}))^T$. Our model for these processes is therefore

$$\begin{bmatrix} \mathbf{S} \\ \mathbf{Z} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0}_n \\ \mathbf{0}_{n_1} \end{bmatrix}, \begin{bmatrix} \Sigma_S & \Sigma_{SZ} \\ \Sigma_{SZ}^T & \Sigma_Z \end{bmatrix} \right), \tag{3}$$

The matrices Σ_S $(n \times n)$, Σ_Z $(n_1 \times n_1)$ are standard covariance matrices for Gaussian processes, specified by some parametric form. The cross-covariance matrix Σ_{SZ} $(n \times n_1)$ accounts for the relationship between the two processes S and Z. Suppose we assume an exponential covariance function, which is a member of the Matérn family of covariance functions, to describe the spatial dependence. Then the (i, j)th elements of our covariance matrices are

$$\begin{aligned} (\Sigma_S)_{ij} = \mathrm{Cov}(S(x_i), S(x_j)) &= \sigma_S^2 \exp(-\theta_S \|x_i - x_j\|), \\ (\Sigma_Z)_{ij} = \mathrm{Cov}(Z(x_i), Z(x_j)) &= \sigma_Z^2 \exp(-\theta_Z \|x_i - x_j\|) \end{aligned} \tag{4}$$

for covariance parameters $\sigma_S^2, \sigma_Z^2, \theta_S, \theta_Z > 0$. We have assumed here, for simplicity, that the covariance is isotropic, that it is only a function of the Euclidean distance between the two locations. The cross-covariance function is constructed as described in Oliver (2003) by taking $\Sigma_{SZ} = \rho_{SZ} L_S L_Z^T$ where the scalar ρ_{SZ} is the correlation between the S and Z processes at the same location, and L_S, L_Z are the Choleski factors of Σ_S, Σ_Z respectively. That is, $\Sigma_S = L_S L_S^T$ and $\Sigma_Z = L_Z L_Z^T$ and

$$(\Sigma_{SZ})_{ij} = \rho_{SZ} (L_S L_Z^T)_{ij}. \tag{5}$$

The matrix in (5) is actually constructed as follows. We first set up a complete

$(n \times n)$ covariance matrix for Z based on *all* locations, not just the ones with $U = 1$. We set up the cross-covariance by taking the product of ρ and the two Choleski factors. We then drop the last $n - n_1$ columns because these are cross-covariances between S and Z in locations where $U = 0$. Hence there is no information regarding Z in these locations and they do not contribute to the likelihood for the observed data.

Oliver (2003) provides details regarding the validity of the above approach for constructing cross-covariances. There are several advantages to using this approach. It allows for greater flexibility in the choice of covariance functions while accommodating fairly limited information about the cross-covariance. In many cases the nature of the spatial dependence of each random field is well established, possibly including situations where these do not have the same covariance structure. At the same time, there might be limited knowledge regarding the spatial covariance between the variables of interest, except perhaps their correlation when these are observed in the same location. In the approach used here the only information one needs to construct a valid cross-covariance function are the individual (possibly of different form) covariance functions and the correlation between the two variables. We found this approach to be convenient and useful when deriving the covariance function under the various assumptions of dependence explored later in the manuscript.

We denote the vector of covariance parameters by $\Theta = (\theta_S, \theta_Z, \rho_{SZ}, \sigma_S^2, \sigma_Z^2)$. For any location $x \in D$ (where as before D encompasses the study region), conditional on S and Z, U and V are mutually independent, with distributions

$$\begin{aligned} f_S(U(x) \mid S(x)) =& f_S(U(x) \mid A(x)), \text{ where } A(x) = E(U(X) \mid S(X), \boldsymbol{\alpha}), \\ \text{and } f_Z(V(x) \mid Z(x)) =& f_Z(U(x) \mid B(x)), \text{ where } B(x) = E(V(X) \mid Z(X), \boldsymbol{\beta}), \end{aligned} \tag{6}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are parameters for the model. $U(x)$ and $V(x)$ depend on the underlying Gaussian process only through their respective expected values $A(x)$ and $B(x)$. And

$$\begin{aligned} h_S(A(x)) =& \mathbf{d}_S(x)\boldsymbol{\alpha} + S(x), \\ h_Z(B(x)) =& \mathbf{d}_Z(x)\boldsymbol{\beta} + Z(x), \end{aligned} \tag{7}$$

where h_S and h_Z are known link functions and $\mathbf{d}_S(x)$ and $\mathbf{d}_Z(x)$ are vectors of explanatory variables and $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are the respective coefficients. The parameters of our model are therefore $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \Theta)$.

We now consider specifics for zero-inflated data such as the CPB data discussed in Section 1. First, we set h_S, h_Z in (7) to the logistic and log-link functions respectively. Then, following the general definitions in (6), we specify

$$\begin{aligned} U(x) &= \text{Bernoulli}(A(x)), \quad \text{so } Pr(U(x) = 1 \mid S(x), \boldsymbol{\alpha}) = A(x) \\ V(x) &= \text{TruncPoisson}(B(x)), \quad \text{so } E(V(x) \mid Z(x), \boldsymbol{\beta}) = \frac{B(x)}{1 - e^{-B(x)}}, \end{aligned} \tag{8}$$

where TruncPoisson is a truncated Poisson random variable (cf. David and Johnson (1952); Plackett (1953)), and hence no zero-valued observations are possible. That is,

$$Pr(V(x) = r \mid B(x)) = \frac{B(x)^r e^{-B(x)}}{r!(1 - e^{-B(x)})}, \;\; r = 1, 2, \ldots$$

To complete the specification of a Bayesian two-stage spatial model, we impose priors on the parameters of the model. We use log-uniform proper priors for the covariance parameters θ_S, θ_Z, and uniform proper priors for $\rho_{SZ}, \sigma_S, \sigma_Z$. The log-uniform prior on a finite interval $\pi(\theta) \propto \theta^{-1}, \log(\theta) \in [t_1, t_2]$ was used by Christensen et al. (2000). Following common practice, we use flat priors for the regression coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}$.

2.4 Model features

The model as described has several desirable features. A two-stage model allows us to examine the features of each component of a mixed response, permitting a closer look at one or both parts as appropriate. The model permits the sets of covariates and fixed effects to differ between the two components, thus allowing the covariates to impact each part of the response in a different way. For instance, the factors determining where CPB large larvae are likely to be found (where the adults have laid eggs), $\mathbf{d}_S(x)$, may not be the same conditions that determine whether they will thrive (i.e., where more of them have survived), $\mathbf{d}_Z(x)$. Even if the covariates are common to both parts, the magnitude of effects, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, may still differ. By embedding the underlying Gaussian process into a more generalized error structure, we expand the class of models that can be modeled directly. Finally, the cross-covariance function Σ_{SZ} allows the two parts of the model to be related. In the CPB example, the strength of the cross-correlation between S and Z relates the severity of infection in location x_i, $V(x_i)$, to incidence in another location x_j, $U(x_j)$ via (5).

3 Sample-based inference

In this section we outline strategies for inference and prediction based on this model, providing details regarding the MCMC algorithms used to explore the posterior distributions of interest.

We have observation vectors $\mathbf{U} = (U(x_1), \ldots, U(x_n))$ and $\mathbf{V} = (V(x_1), \ldots, V(x_{n_1}))$, whose mean values depend on the regression parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as well as the unobserved and potentially related Gaussian processes $\mathbf{S} = (S(x_1), \ldots, S(x_n))$ and $\mathbf{Z} = (Z(x_1), \ldots, Z(x_{n_1}))$. As discussed in Subsection 2.3, $x_1, \ldots, x_n$ are locations in D and $n_1 \leq n$. We can summarize this model as follows:

- Stage 1: The observation vectors $\mathbf{U}$ and $\mathbf{V}$ are modeled via logistic and log links respectively, conditional on parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and the underlying spatial processes $\mathbf{S}, \mathbf{Z}$ as described in (7) and (8).
- Stage 2: $\mathbf{S}, \mathbf{Z}$ are jointly modeled via a zero-mean Gaussian process with covariances as described in (3),(4), and (5).
- Stage 3: Priors for $\boldsymbol{\alpha}, \boldsymbol{\beta}, \Theta$ are specified as in Subsection 2.3.

The above stages can be combined to derive the posterior distribution of the parameters given the observations, $\pi(\mathbf{S}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \Theta \mid \mathbf{U}, \mathbf{V})$. This is the distribution on which inference regarding the parameters is based. Since this distribution is analytically intractable, we rely on MCMC algorithms to perform sample-based inference. This distribution is high dimensional and the strong dependence among components makes it challenging to construct good MCMC algorithms to simulate from it. Furthermore, unlike in the areal data case where Gaussian Markov random field modeling implies a conditional independence structure which circumvents the need for matrix computations (cf. Agarwal et al. (2002); Ver Hoef and Jansen (2007)), each iteration of our MCMC algorithm is expensive due to the calculations involving dense covariance matrices. Constructing a fast mixing sampler is therefore even more critical in our case. We develop an MCMC algorithm based on a version of the Langevin-Hastings as described in Christensen et al. (2006). We provide details on how we construct our MCMC sampler in the appendix.

To ensure that our MCMC based estimates were reliable, we used standard heuristics such as starting the chain from different initial values and comparing resulting estimates. To determine how long to run the Markov chains, we used a stopping rule based on Monte Carlo standard errors for the posterior mean estimates computed by consistent batch means (Flegal et al., 2008; Jones et al., 2006): when the standard errors for all parameter estimates were low enough, the chain

was stopped. For instance, in the MCMC algorithm for the TSF model, the Monte Carlo standard errors of the posterior mean estimates of all regression parameters and ρ_{SZ} were 0.01 or smaller. The MCMC algorithms were implemented in `R` (Ihaka and Gentleman, 1996). Inference was computationally intensive. For instance, for the simulated data example involving 400 locations, it takes around 60 hours to complete 100,000 full iterations of the MCMC algorithm on a 3.0 GHz quadcore Intel Xeon processor with 32 gigabytes of memory.

Our sample-based procedure is a powerful approach for obtaining predictions of the incidence and prevalence processes at locations where there are no observations, and it also provides a convenient way to obtain estimates of the underlying smooth latent spatial fields and any other functions of these processes that may be of interest. However, one must be very cautious when interpreting the parameters of the spatial regressions in these models. Due to identifiability issues, the covariance parameter estimates as well as the regression parameter estimates may be suitable for a predictive model, but may not be easy to interpret, as we will later describe in the context of our simulation study in Section 4. We note that the issues we outline here are not unique to our model or even to zero-inflated spatial data. They are common to spatial generalized linear mixed models (SGLMMs), as has been pointed out as early as in the original paper that describes the framework for spatial generalized linear mixed models (Diggle et al., 1998). The confounding between the spatial random effects (the latent Gaussian process used to incorporate dependence) and the fixed effects (the regression parameters) has, more recently, been noted and studied by others including Reich et al. (2006) and Paciorek (2009). In addition, Zhang (2002, 2004) establishes both via theory and simulation that learning about covariance function parameters can be difficult due to confounding (inconsistent estimators in the maximum likelihood context), both for linear spatial models as well as SGLMMs. The regression parameter and covariance parameter identifiability is-

sues are not unique to our modeling approach, but we point them out here and later in our simulation study since we believe that while these models are very useful for prediction, their parameters should not be over-interpreted.

4 Application to data examples

We now study our modeling and computational approach in the context of both simulated and real data. In studying the two-stage model, we also considered a few different covariance structures that are special cases of our model. The first covariance structure is as in the *full two-stage* (TSF) model described in Section 2, with dependence among random effects for counts ($\mathbf{Z}$), dependence among random effects for the binaries ($\mathbf{S}$), and cross-correlation among $\mathbf{Z}$ and $\mathbf{S}$. The second is a simpler covariance with dependence among $\mathbf{Z}$ but independence among $\mathbf{S}$, and cross-correlation among $\mathbf{Z}$ and $\mathbf{S}$, henceforth the *two-stage independent binary* (TSIB) model. The third, which we call the *two-stage no correlation* (TSNC) model, takes the TSF model but removes the cross-correlation among $\mathbf{Z}$ and $\mathbf{S}$. We note that this model is analogous to the hurdle model studied in Ver Hoef and Jansen (2007) in the context of areal data. Finally, the fourth covariance structure we studied assumes dependence among $\mathbf{Z}$, independence among $\mathbf{S}$, and no cross-correlation among $\mathbf{Z}$ and $\mathbf{S}$. We refer to this as the *two-stage independent binary no cross-correlation* (TSIBNC) model.

4.1 An application to simulated data

We first describe the application of our model and computational methods to a simulated data set.

4.1.1 Description

We simulated data by generating a two-stage response in 2,601 equally spaced locations over the unit square. In the ith location, say x_i, the two-stage response $(U(x_i), V(x_i))$ was simulated following the model described in Section 2 with

$$\begin{aligned} U(x_i)|S(x_i) &\sim \text{Bernoulli}(A(x_i)) \\ V(x_i) \mid Z(x_i), U(x_i) &\sim \text{Truncated Poisson}(B(x_i)) \\ \text{logit}(A(x_i)) &= \alpha_0 + d(x_i)\alpha_1 + S(x_i) \\ \log(Bi) &= \beta_0 + d(x_i)\beta_1 + Z(x_i). \end{aligned}$$

Conditionally on the S and Z observations, for any two locations i and j, the pairs $(U(x_i), V(x_i))$ and $(U(x_j), V(x_j))$ are independent, and the $S(x_i)$ and $Z(x_i)$ are stationary zero-mean processes with covariances following the exponential covariance function $C(\delta_{ij}) = \sigma^2 \exp(-\theta\delta_{ij})$, where δ_{ij} is the Euclidean distance between locations x_i and x_j. The cross-covariance is constructed as described in Section 2. The explanatory variable $d(x_i)$ is a function of location along the horizontal axis, $d(x_i) = 2x_i + (0.01)W_i$ where $W_i \sim N(0, 1)$. The regression parameter values were set to $(\alpha_0, \alpha_1) = (2, 5)$ and $(\beta_0, \beta_1) = (1, 3)$ and the covariance parameters were set to $(\sigma^2, \theta_S, \theta_Z, \rho_{SZ}) = (1, 10, 5, 0.75)$. The regression parameters were chosen to give a substantial proportion of zeros in the sample, yet induce a clear spatial trend in mean counts. For instance, $\beta_1 = 3$ means that under a constant signal (Z_0 say), the expected count (conditional on having observed presence, that is, at least 1 count) increases from 1 to about 20 from the left side of the field to the opposite end. The chosen covariance parameters induced moderate spatial correlation among the S and among the Z, as well as between S and Z. For instance, at $\theta_S = 10$, the correlation between neighboring S signals goes down to about 0.15 at a scaled distance of about 0.2, so that S signals are essentially uncorrelated at distances longer than one- fifth

Table 1: 95% highest posterior density intervals for regression parameters for the **simulated data**. Note: parameters may not be directly comparable since they have different interpretations under the different models.

Parameter	TSF Model	TSIB Model	TSNC Model	TSIBNC Model
$\alpha_0 = 2.0$	(1.95,3.22)	(4.00,8.65)	(1.94,3.22)	(6.42,8.37)
$\alpha_1 = 5.0$	(4.69,7.07)	(9.71,20.27)	(4.71,7.05)	(13.41,17.36)
$\beta_0 = 1.0$	(1.56,1.78)	(1.51,1.74)	(1.56,1.79)	(1.56,1.79)
$\beta_1 = 3.0$	(0.78,1.18)	(0.84,1.24)	(0.78,1.19)	(0.78,1.18)

of the field. The correlation between S and Z at the same location is $\rho_{SZ} = 0.75$, and thereafter decays exponentially with distance. Figure 3 is a plot of the simulated observations observed at 400 sample locations. There were 127 locations with zero incidence.

— INSERT FIGURE 3 ABOUT HERE —

4.1.2 Results

We consider two aspects of the performance of our models for our simulated data set: prediction of the spatial process at unobserved locations and inference for the model parameters. Table 1 provides estimated 95% highest posterior densities (HPD) credible regions of the regression parameters using the approximate procedure of Chen et al. (2000). The regression parameters for incidence (α_0, α_1) are captured well in the TSF and TSNC models but the intervals for the TSIB and TSIBNC models, which ignore spatial dependence among incidences, do not capture the true values well. None of the models captures well the true regression parameters for prevalence (β_0, β_1). We suspect that this is at least partly due to the simulated random effects $(\mathbf{Z})$ exhibiting a decreasing trend along the x-axis, which directly counters the increasing trend in mean that we had imposed on the model. We note that Diggle et al. (1998) report similar findings for regression parameters in simulated

spatial count data and attribute it to the fact that the regression parameters need to be interpreted conditional on the dependent random effects (see also Diggle et al., 1994). Although Agarwal et al. (2002); Rathbun and Fei (2006); Ver Hoef and Jansen (2007) do not report results from applications to simulated data in the context of zero-inflated spatial data, we believe that one would likely obtain similar results for regression parameters in their models. Identifying individual covariance parameters is particularly challenging so we fix σ_S, σ_Z at 1 in order to identify the remaining parameters. For a discussion of related identifiability issues in spatial models for binary data and SGLMMs more generally, see De Oliveira (2000); Zhang (2002, 2004). For both incidence and prevalence, we are able to infer spatial dependence even though our estimated covariance parameters may not always agree with the the true values used in the simulation study. Notably, the cross-correlation between the incidence and prevalence random effects ($\mathbf{S}$ and $\mathbf{Z}$ processes) was not captured, that is, the 95% HPD of ρ_{SZ} included 0 in spite of the strong cross-correlation in the simulated data.

Prediction is arguably the most important criteria for assessing the performance of these models. We find that all four of the models we study produce very similar predictions for the zero-inflated random variable ($\mathbf{Y}$). For instance, Figure 4 illustrates the predictions of $\mathbf{Y}$ based on the TSNC model; $\mathbf{Y}$ predictions for all other models are virtually identical. This suggests that the simplest model, TSIBNC, may be adequate when computational considerations are critical and prediction of $\mathbf{Y}$ is the only goal. The predictions of the prevalence process $\mathbf{V}$ are also similar across the models as displayed in Figures 6 and 7 for the TSNC and TSIB models respectively. On the other hand, the models that ignore spatial dependence among the binary values $\mathbf{U}$ produce predictions of incidences that lack smoothness, as can be seen by contrasting the predictions of $\mathbf{U}$ for the TSNC and the TSIB models in Figures 5 and 7 respectively. The TSF model produced similarly smooth predictions

of **U** while the TSIBNC model produced predictions similar to the TSIB model. Hence, the TSF and TSNC models are superior to the TSIB and TSIBNC models since they produce predictions that are consistent with the underlying assumption of smoothness in both incidence and prevalence processes. As discussed above, the TSF and TSNC models will likely also provide reasonable estimates of the regression parameters for incidence and therefore may be preferable to the other models. Given the fact that computation for the TSNC model is faster than for the TSF model and that the cross-correlation (ρ_{SZ}) is hard to infer, our recommendation for such problems is therefore the TSNC model.

In addition to the above simulated data set, we also simulated another data set that resembled more closely the Colorado Potato Beetles (CPB) data set analyzed in Subsection 4.2. In particular, our simulated data set used a sampling design identical to the one used in the real data, and we used parameter values that resulted in a data set with similar characteristics to the real data set. An important reason for conducting this additional study was to find out whether the particular sampling design used in the CPB analysis would affect the conclusions, say by introducing edge effects or row effects. For brevity, we do not include a detailed discussion of the results of our study here, but the conclusions are qualitatively the same as above. In this simulated data example we find again that we prefer the TSNC model to the others for reasons similar to those described above. Also, as discussed above and in Section 3, we conclude that one must exercise caution in the interpretation of regression parameters.

— INSERT FIGURE 4 ABOUT HERE —

— INSERT FIGURE 5 ABOUT HERE —

— INSERT FIGURE 6 ABOUT HERE —

— INSERT FIGURE 7 ABOUT HERE —

4.2 An application to ecology

We now describe the application of our models and computing algorithms to a data set on Colorado Potato Beetles, our motivating example from Section 1.

4.2.1 Description

In the second application, we revisit the entomological study in which different life stages of Colorado potato beetle were counted weekly at a resolution of one meter-row. The data set considered here consists of large larvae count taken at week eight. There were 296 observations taken in a systematic sampling pattern in an 80-m square field. Figure 1 shows the observations in the sampled locations. The 296 observations consist of 144 zeros and 152 positive counts. Each observation at location x_i, $Y(x_i)$ was transformed into a two-stage response $(U(x_i), V(x_i))$ as before, and we fit the same set of two-stage models discussed previously — the TSF, TSIB, TSNC and TSIBNC models. Due to the location and orientation of the experimental plot, it is believed that the source of infestation (immigrating adults) would be the north side of the field. Therefore, $d(x_i)$ is taken as scaled and centered northing coordinate, the single explanatory variable in the simple mean functions of Section 2.3. We consider the same model structures as we did in Section 4.1. In generating posterior samples of $\mathbf{S}$ and $\mathbf{Z}$ following the algorithm described in Section A.2, we used truncation constant $K = 50$ for $\bigtriangledown(\gamma)^{trunc}$ and variance scale parameter $k = 0.40$. These scaling parameters were selected via trial and error. After a few short preliminary runs for different values, we could tell which values resulted in more efficient algorithms by simply examining either the MCMC standard errors for estimates of expected values of different parameters of the posterior distributions (smaller standard errors obviously reflecting less autocorrelation), or by simply looking at the autocorrelation plots of the samples.

Table 2: 95% highest posterior density intervals for regression parameters for the **CPB data**. Note: parameters may not be directly comparable since they have different interpretations under the different models.

Parameter	TSF Model	TSIB Model	TSNC Model	TSIBNC Model
α_0	(-0.31,0.41)	(-0.47,0.91)	(-0.24,0.42)	(-1.08,2.59)
α_1	(1.81,4.29)	(4.05,9.99)	(1.94,4.23)	(7.84,15.71)
β_0	(1.44,1.72)	(1.45,1.73)	(1.47,1.75)	(1.44,1.74)
β_1	(0.71,1.46)	(0.80,1.57)	(0.73,1.50)	(0.70,1.46)

4.2.2 Results

Table 2 summarizes posterior estimates for the regression parameters in all four models. As in the simulated data set, the cross-correlation parameter (in models TSF and TSIB) was not found to be significant (the 95% HPD included 0). The positive mean values of both slopes (α_1, β_1) are consistent with the expectation that locations further to the north (higher along the y-axis) have higher densities of large larvae because the source of infestation is just north of this field. For instance, given a constant S and taking 1.90 (from Table 2) as our estimate of α_1, the odds of finding at least one large larva increases from 1 in the middle of the field to about $e^{(0.5)(1.9)} = 2.6$ to the north end of the field.

For reasons discussed in Section 4.1, we prefer the TSF and TSNC models — although predictions for the zero-inflated and count processes $(\mathbf{Y}, \mathbf{V})$ are fairly similar across all four models, the TSF and TSNC models produce smoother predictions for the incidence $(\mathbf{U})$ process. Means for the posterior predictive distributions for the large larvae counts are shown in Figure 8. There is a clear increasing trend in the predicted mean as we move closer to the north edge of the field, and an increase in variability as well. In addition to the mean, we can also map other functionals of interest. For instance, to identify possible 'hot spots' or areas that may need control measures, we can compute the rate at which each location has a predicted mean in the upper (say, 10%) quantile, or has a mean that exceeds a known threshold.

Figure 9 shows mean predicted incidence (E(U)) and positive counts (E(V)). The map of mean predicted incidence (E(U)) shows only a generally increasing mean trend along the y-axis, and some localized variation where positive counts were observed in the sampled locations. The mean predicted positive counts (E(V)) shows a similar general increasing trend along the y-axis. It is possible that the spatial processes that affect incidence may be different from those that drive prevalence. Blom and Fleischer (2001) found that the distribution of adults followed a mean trend, with higher densities observed closer to sources of immigrating adults. However, they observed little or no spatial dependence. This may mean that the adults, once they are in the field, have no preference for particular locations or conditions to lay their eggs. Therefore, it may turn out that where the eggs (and therefore the larvae) are found will also exhibit no spatial correlation. However, non-uniform conditions within the potato field may determine how many of these eggs will survive to become large larvae, which could explain why some spatial dependence can be observed among large larvae. Thus, it may be that incidence and prevalence are really two different processes with different covariance structures.

5 Discussion

We have described an approach for modeling zero-inflated point level spatial count data. Our model is particularly useful for situations where the underlying data generating mechanism suggests separate but dependent processes for incidence and prevalence. Our SGLMM framework allows us to incorporate spatial dependence and cross-correlations among the incidence and prevalence processes. Our study of various versions of our two-stage spatial model suggests that if the goal is only prediction of the zero-inflated counts, our simplest two-stage model (TSIBNC) without cross-correlations or spatial dependence among incidences, works well. However,

in many situations, including our motivating example from ecology, predicting and understanding the smoothed binary incidence process is also important. In such cases, models that incorporate dependence both among the incidences and among the prevalences (TSF and TSNC models) are superior to models that assume the incidences are independent (TSIB and TSIBNC models). The TSF and TSNC models also offer some advantages in terms of providing more easily interpretable regression parameters when modeling incidence. We also found that inferring the cross correlation among the incidence and prevalence spatial processes (ρ in the TSF and TSIB models) seems to be very challenging. Given our goals for prediction and inference, we therefore recommend the TSNC model, a two-stage model that assumes the incidences are dependent, the prevalences are dependent, but that the incidences and prevalences are not cross-correlated. We believe our methods provide a very sound approach to spatial prediction, and we can use our fitted model to predict incidence and prevalence while incorporating spatial dependence. However, as discussed in Sections 3 and 4, the covariance parameters and regression coefficients in spatial generalized linear models should not be over-interpreted. Our Langevin-Hastings based MCMC algorithms produce well-mixing Markov chains for several different models and both simulated and real data. As data sets get larger, the computationally time for the Langevin-Hastings algorithm may become prohibitive. In such cases, it would be of interest to investigate new, potentially more efficient versions of the algorithm. For instance, Dostert et al. (2006) and Efendiev et al. (2006) use coarse-scale models to compute the necessary gradients, and Girolami and Calderhead (2011) propose a generalized version of the Langevin-Hastings algorithm based on Riemann geometry. Exploring such algorithms in conjunction with recent approaches for modeling large spatial data sets (cf. Banerjee et al., 2008; Cornford et al., 2005; Cressie and Johanneson, 2008; Higdon, 1998) may also be a fruitful avenue for future research.

Acknowledgments

The authors are grateful to Shelby Fleischer and Paul Blom for the data set, and to the associate editor and two anonymous referees for very helpful comments and suggestions.

A MCMC for the two-stage model

A.1 Outline of MCMC-based approach

All marginal distributions are available using samples from $\pi(\mathbf{S}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \Theta \mid \mathbf{U}, \mathbf{V})$, so we can easily infer the dependence and error parameters (Θ) along with variability in our estimates. We can use the posterior distribution of the regression parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to test the significance of individual coefficients, and thereby study the importance of individual predictors. It is important to note that these parameters have a conditional interpretation, in that $\boldsymbol{\alpha}$ reflects the effect of the covariates $\mathbf{d}_S(x)$ on $E[U(x) \mid S(x)]$ and $\boldsymbol{\beta}$ the effect of covariates $\mathbf{d}_Z(x)$ on $E[V(x) \mid Z(x), U(x) = 1]$ (see also Diggle et al., 1998). Reich et al. (2006) describe how inference for the regression parameters can be substantially affected by spurious collinearity between the covariate and the spatial random effects.

Suppose we are given a set of m new locations at which no observations are available, say $x_1^*, \ldots, x_m^*$, and we are interested in estimating U and V at these locations. This can be done by first inferring the S and Z processes at these locations, that is, $\mathbf{S}^* = (S(x_1^*), \ldots, S(x_m^*))$ and $\mathbf{Z}^* = (Z(x_1^*), \ldots, Z(x_m^*))$, and then inferring U and V based on $\mathbf{S}^*, \mathbf{Z}^*$. Once we have samples from the joint posterior distribution $\pi(\mathbf{S}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \Theta \mid \mathbf{U}, \mathbf{V})$, we can infer $\mathbf{S}^*, \mathbf{Z}^*$ by sampling from the posterior predictive distribution as follows:

- Sample from the posterior distribution of $\pi(\mathbf{S}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \Theta \mid \mathbf{U}, \mathbf{V})$ via Markov

chain Monte Carlo. Details are provided in Subsection A.2.

- Given a sampled vector of $(\mathbf{S}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \Theta)$ from above, we can easily sample the vector $(\mathbf{S}^*, \mathbf{Z}^*)$ from the posterior predictive distribution as it is simply a multivariate normal density given $(\mathbf{S}, \mathbf{Z})$.

To infer the U, V process at these new locations, we can utilize the S, Z draws from above along with corresponding $\boldsymbol{\alpha}, \boldsymbol{\beta}$ samples. Predictions about the Y process at unobserved locations are easily obtained in similar fashion from the U, V posterior predictive draws. Finally, other quantities of interest may be easily obtained from the sampled values of $\mathbf{U}, \mathbf{V}, \mathbf{Y}$. For instance, upper quantiles of incidence will show where the highest risk for incidence lies. The probability that mean count will exceed a given threshold will also reveal areas that potentially require some management intervention. Since our approach relies heavily on estimates based on MCMC, developing efficient MCMC algorithms is an important part of this work.

A.2 Markov chain Monte Carlo implementation details

The simplest default MCMC algorithm for this model would involve univariate updates where the covariance parameters, the regression parameters, and each of the random effects are updated in turn in each iteration of the MCMC algorithm. Unfortunately, the Gaussian processes used for modeling random effects in point-level data do not lend themselves to this relatively simple univariate MCMC algorithm. We therefore pursued the following approach. Each of the covariance and regression parameters was updated using univariate Metropolis updates — normal proposals centered at the current value of the parameter. For the random effects vector, this method of updating is computationally intensive since each update of each random effect $S(x_i)$ or $Z(x_j)$ involves matrix computations of an $(n + n_1) \times (n + n_1)$ dimensional matrix, with the number of floating point operations typically of order

$(n + n_1)^3$ at each iteration. Worse yet, such a scheme produces a very slow mixing Markov chain that does not explore the posterior distribution efficiently, resulting in poor estimates of the posterior distributions. As is well known, block updating schemes can, in principle, greatly improve mixing (Liu et al., 1994) while simultaneously reducing the number of expensive matrix computations. However, constructing proposals for blocks of highly dependent parameters in spatial models can be challenging (cf. Christensen et al., 2006; Christensen and Waagepetersen, 2002; Haran et al., 2003; Knorr-Held and Rue, 2002). Christensen and Waagepetersen (2002) propose a Langevin-Hastings MCMC (LHMCMC) algorithm for spatial count data which simultaneously updates the entire vector of random effects or regression coefficients based on gradient information, and show that, in some cases, the resulting algorithm is provably fast mixing (geometrically ergodic) and fairly efficient in practice. We outline this approach below.

Let $\Sigma^{1/2}$ be the Choleski factor of the covariance of (S, Z) and let

$$\bigtriangledown(\gamma) = \frac{\partial}{\partial \gamma} \log f(\gamma \mid \dots) = -\gamma + (\Sigma^{1/2})^T \begin{bmatrix} \left\{ (U(x_i) - A(x_i)) \frac{h_c^{'}(A(x_i))}{h^{'}(A(x_i))} \right\}_{i=1}^{n} \\ \left\{ (V(x_j) - B(x_j)) \frac{g_c^{'}(B(x_j))}{g^{'}(B(x_j))} \right\}_{j=n+1}^{n+n_1} \end{bmatrix}$$

denote the gradient of the log target density (denoted by $f(\gamma \mid \dots)$) where $h_c^{'}$ and $g_c^{'}$ are the partial derivatives of the canonical functions for the Binomial and Poisson distributions, respectively, and $h^{'}$ and $g^{'}$ are partial derivatives of the actual link functions we used for the application. As before, $A(x_i)$ and $B(x_i)$ are the means for $U(x_i)$ and $V(x_i)$ conditional on $(S(x_i), \boldsymbol{\alpha})$ and $(Z(x_i), \boldsymbol{\beta})$ respectively. Since we used canonical links in both cases, $\frac{h_c^{'}(A(x_i))}{h^{'}(A(x_i))} = \frac{g_c^{'}(B(x_i))}{g^{'}(B(x_i))} = 1$ for each i. Also, for the truncated Poisson GLMM proposed here, Christensen and Waagepetersen (2002) have shown that the LHMCMC algorithm is not geometrically ergodic because $\bigtriangledown(\gamma)$ increases very fast when γ approaches infinity in some directions. Hence, our

truncated gradient is,

$$\bigtriangledown^{trunc}(\gamma) = \frac{\partial}{\partial\gamma}\log f(\gamma \mid \dots) = -\gamma + (\Sigma^{1/2})^T \begin{bmatrix} \{U(x_i) - A(x_i)\}_{i=1}^{n} \\ \{V(x_j) - (B(x_j) \wedge K)\}_{j=n+1}^{n+n_1} \end{bmatrix}, \quad (9)$$

where $K \in (0, \infty)$ is a truncation constant. This results in a geometrically ergodic LHMCMC algorithm. The binomial part of the gradient does not need to be truncated because the mean $(A(x_i))$ is bounded.

The Langevin-Hastings update simply involves using a multivariate normal proposal with mean vector $\xi(\gamma) = \gamma + \frac{k}{2} \bigtriangledown (\gamma)^{trunc}$ and covariance matrix $kI, k > 0$. The main advantage of using the Langevin-Hastings update is that it simultaneously updates the entire vector of random effects based on gradient information and can be more efficient. However, in implementing the LH algorithm above we encountered problems with mixing. We therefore utilized the modifications to the Langevin-Hastings algorithm for SGLMMs as suggested by Christensen et al. (2006). Roberts and Rosenthal (2001) observed that the LH algorithm is sensitive to inhomogeneity of the components; it loses efficiency when components have different variances. Christensen et al. (2006) showed that this can arise in SGLMMs because the variability of individual components of the target density can vary depending on the observation at each location. For instance, for Poisson observations with a log link they showed that large observations tend to be more informative about their mean than small ones are, so that generally the variance of $S(x_i)|Y(x_i)$ will be smaller in locations with relatively high counts. Conversely, the variance of $S(x_i)|Y(x_i)$ will generally be higher in locations with smaller counts. Therefore, locations with higher counts (smaller variance) will tend to reject more proposals, while moves will generally be smaller than optimal for components with large variance (lower counts). Overall, total mixing of S will be slower than if variances were equal. In the bino-

mial case, the variance increases when the observed value approaches 0 or $m(x_i)$, the number of trials at location x_i. For binary spatial random effects the variance for $S(x_i) \mid Y(x_i)$ is uniformly high for all locations.

To improve the mixing of the Markov chain in the presence of inhomogeneity and highly correlated components, we follow Christensen et al. (2006) and transform the vector of random effects into *a posteriori* uncorrelated components with homogeneous variance. The covariance matrix for $S \mid y$ is approximately $\tilde{\Sigma} = (\Sigma^{-1} + \Lambda(\hat{\mathbf{S}}))^{-1}$ where $\Lambda(\hat{\mathbf{S}})$ is a diagonal matrix with entries $\frac{\partial^2}{(\partial S(x_i))^2} \log f(Y(x_i) \mid S(x_i)), i = 1, \ldots, n$, and $\hat{\mathbf{S}}$ is a typical value of $\mathbf{S}$, such as the posterior mode of $\mathbf{S}$. Let $\tilde{\mathbf{S}}$ be such that $\mathbf{S} = \tilde{\Sigma}^{1/2}\tilde{\mathbf{S}}$. $\tilde{\mathbf{S}}$ therefore has approximately uncorrelated components with homogeneous variance, simplifying the construction of an efficient MCMC algorithm. For our application, setting $\Lambda(S(x)) = 0$ for all x appears to be adequate, though other possibilities discussed in Christensen et al. (2006) can also be explored. We also note that other versions of the LHMCMC algorithm may be worth exploring as well, including adaptive versions (e.g. Atchade, 2006; Marshall and Roberts, 2009), as well as the geometric approach in Girolami and Calderhead (2011).

References

Agarwal, D. K., Gelfand, A. E., and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, 9(4):341–355.

Atchade, Y. (2006). An Adaptive Version for the Metropolis Adjusted Langevin Algorithm with a Truncated Drift. *Methodology and Computing in Applied Probability*, 8:235–254.

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall Ltd.

Banerjee, S., Gelfand, A., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 70:825–848.

Blom, P. and Fleischer, S. (2001). Dynamics in the Spatial Structure of Leptinotarsa decemlineata (Coleoptera: Chrysomelidae). *Environmental Entomology*, 30(2):350–364.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag Inc.

Christensen, O., Møller, J., and Waagepetersen, R. (2000). Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte carlo. Technical report, Aalborg University, Department of Mathematical Sciences.

Christensen, O. F., Roberts, G. O., and Sköld, M. (2006). Robust Markov Chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):1–17.

Christensen, O. F. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, 58(2):280–286.

Cornford, D., Csato, L., and Opper, M. (2005). Sequential, Bayesian Geostatistics: A Principled Method for Large Data Sets. *Geographical Analysis*, 37(2):183–199.

Cressie, N. and Johanneson, G. (2008). Fixed rank kriging for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 70:209–226.

David, F. and Johnson, N. (1952). The truncated Poisson. *Biometrics*, 8(4):275–285.

De Oliveira, V. (2000). Bayesian prediction of clipped Gaussian random fields. *Computational Statistics & Data Analysis*, 34(3):299–314.

Diggle, P., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press.

Diggle, P. and Ribeiro, P. (2007). *Model-based Geostatistics*. Springer-Verlag Inc.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics (Disc: P326-350). *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 47:299–326.

Dostert, P., Efendiev, Y., Hou, T., and Luo, W. (2006). Coarse-gradient Langevin algorithms for dynamic data integration and uncertainty quantification. *Journal of Computational Physics*, 217:123–142.

Efendiev, Y., Hou, T., and Luo, W. (2006). Preconditioning Markov chain Monte Carlo simulations using coarse-scale models. *SIAM Journal on Scientific Computing*, 28:776–803.

Fernandes, M., Schmidt, A., and Migon, H. (2009). Modelling zero-inflated spatio-temporal processes. *Statistical Modeling*, 9:3–25.

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

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B*, 73:1–37.

Gschlößl, S. and Czado, C. (2008). Modelling Count Data with Overdispersion and Spatial Effects. *Statistical Papers*, 49:531–552.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, 35:403–410.

Haran, M. (2011). Gaussian random field models for spatial data. In *Handbook of Markov chain Monte Carlo, Eds. Brooks,S.R., Gelman, Andrew, Jones, G.L. and Meng, X.L.* (to appear) . Chapman and Hall/CRC.

Haran, M., Hodges, J. S., and Carlin, B. P. (2003). Accelerating computation in Markov random field models for spatial data via structured MCMC. *Journal of Computational and Graphical Statistics*, 12:249–264.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean (Disc: P191-192). *Environmental and Ecological Statistics*, 5:173–190.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov Chain Monte Carlo. *Journal of the American Statistical Association*, 101(476):1537–1547.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614.

Legendre, P. and Fortin, M. (1989). Spatial pattern and ecological analysis. *Plant Ecology*, 80(2):107–138.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81:27–40.

Marshall, T. and Roberts, G. (2009). An ergodicity result for Adaptive Langevin Algorithms. Technical report, University of Warwick, Centre for Research in Statistical Methodology (CRiSM).

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Chapman & Hall Ltd.

Oliver, D. (2003). Gaussian Cosimulation: Modelling of the Cross-Covariance. *Mathematical Geology*, 35(6):681–698.

Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96(454):730–745.

Paciorek, C. (2009). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Harvard University Biostatistics Working Paper Series*, page 98.

Plackett, R. L. (1953). The truncated poisson distribution. *Biometrics*, 9(4):485–488.

Rathbun, S. L. and Fei, S. (2006). A spatial zero-inflated Poisson regression model for oak regeneration. *Environmental and Ecological Statistics*, 13(4):409–426.

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206.

Roberts, G. and Rosenthal, J. (2001). Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science*, 16(4):351–367.

Rossi, R., Mulla, D., Journel, A., and Franz, E. (1992). Geostatistical Tools for Modeling and Interpreting Ecological Spatial Dependence. *Ecological Monographs*, 62(2):277–314.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall.

Schotzko, D. and O'Keeffe, L. (1989). Geostatistical description of the spatial distribution of lygus hesperus (heteroptera: Miridae) in lentils. *Journal of Economic Entomology*, 82(4):1277–1288.

Schotzko, D. and Smith, C. (1991). Effects of preconditioning host plants on population develoment of Russian wheat aphids(Homoptera: Aphididae). *Journal of Economic Entomology*, 84(3):1083–1087.

Ver Hoef, J. and Jansen, J. (2007). Space–time zero-inflated count models of Harbor seals. *Environmetrics*, 18(7):697.

Williams, L., Schotzko, D., and McCaffrey, J. (1992). Geostatistical Description of the Spatial Distribution of Limonius californicus (Coleoptera: Elateridae) Wireworms in the Northwestern United States, with Comments on Sampling. *Environmental Entomology*, 21(5):983–995.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, 58(1):129–136.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

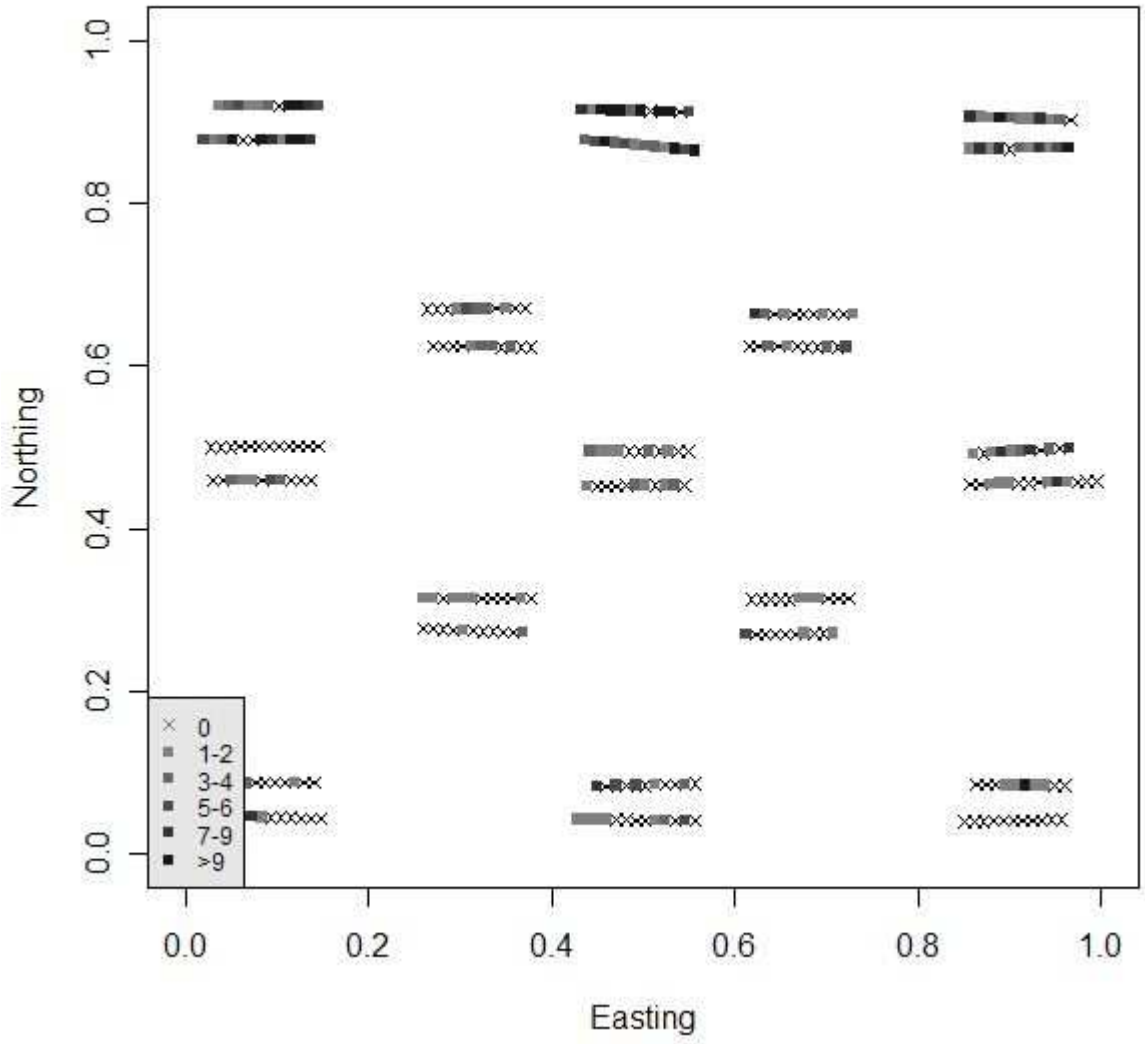

Figure 1: The spatial distribution of the raw counts of Colorado Potato Beatles (CPB), with the x and y-coordinates on the field both transformed to the range (0,1)

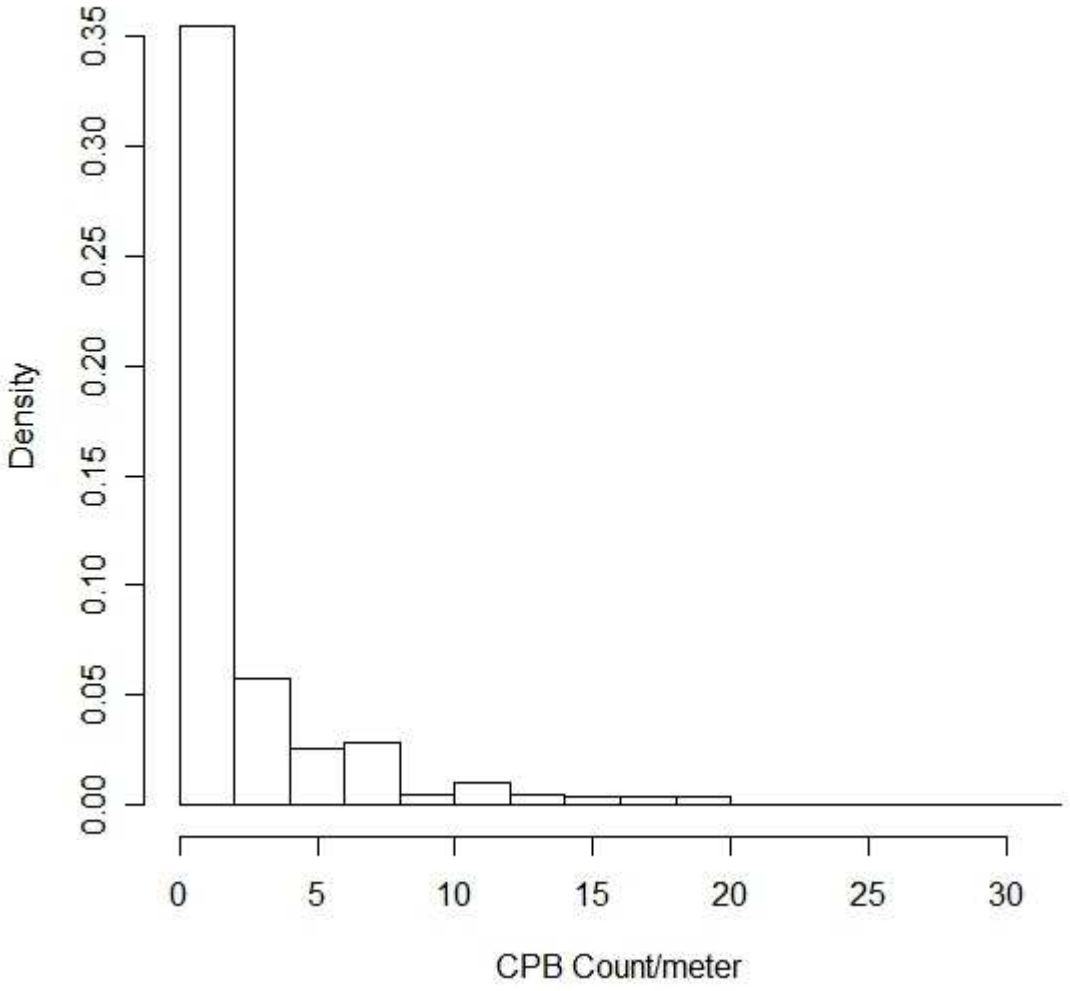

Figure 2: A histogram summary of the raw counts of Colorado Potato Beatles (same as in Figure 1)

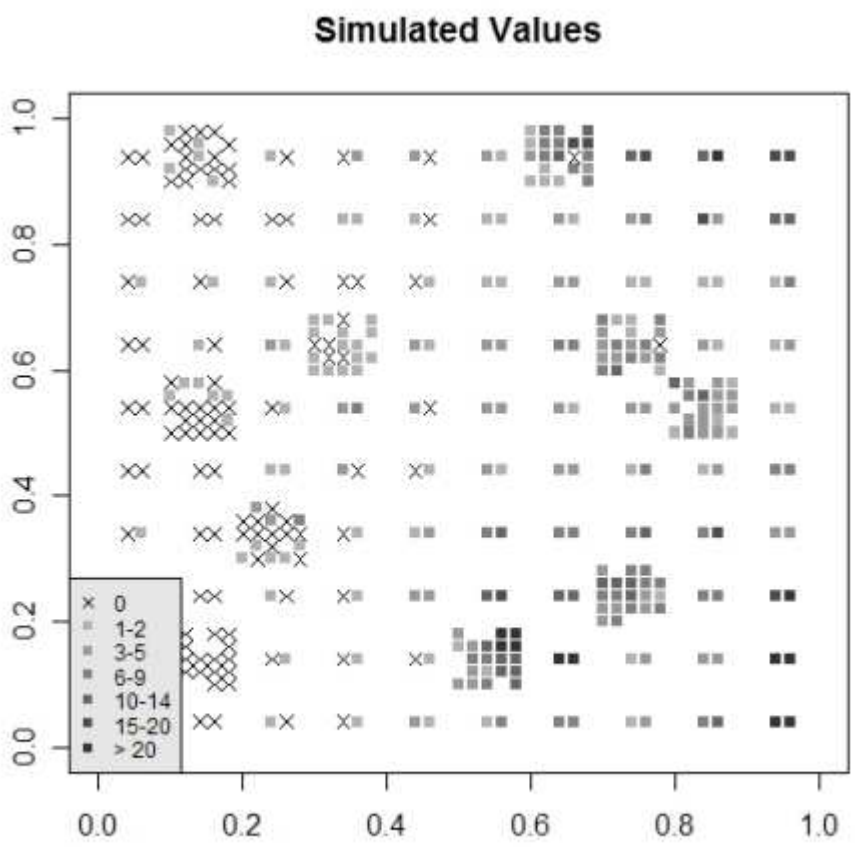

Figure 3: Data simulated from the two-stage ZIP model as described in Section 4.1

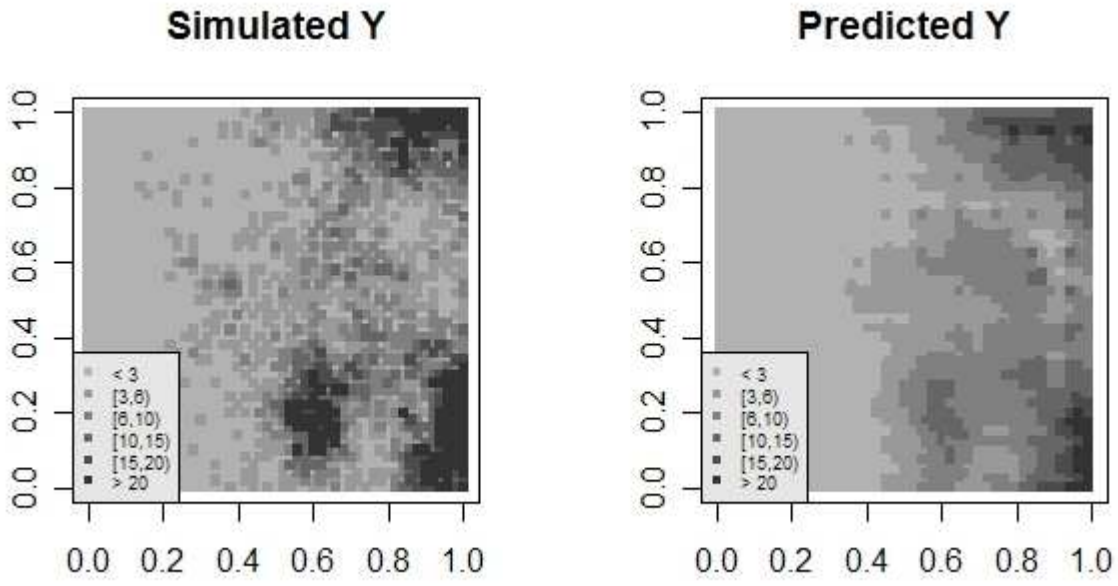

Figure 4: Simulated observations (Y) and predicted observations (posterior means) from TSNC model

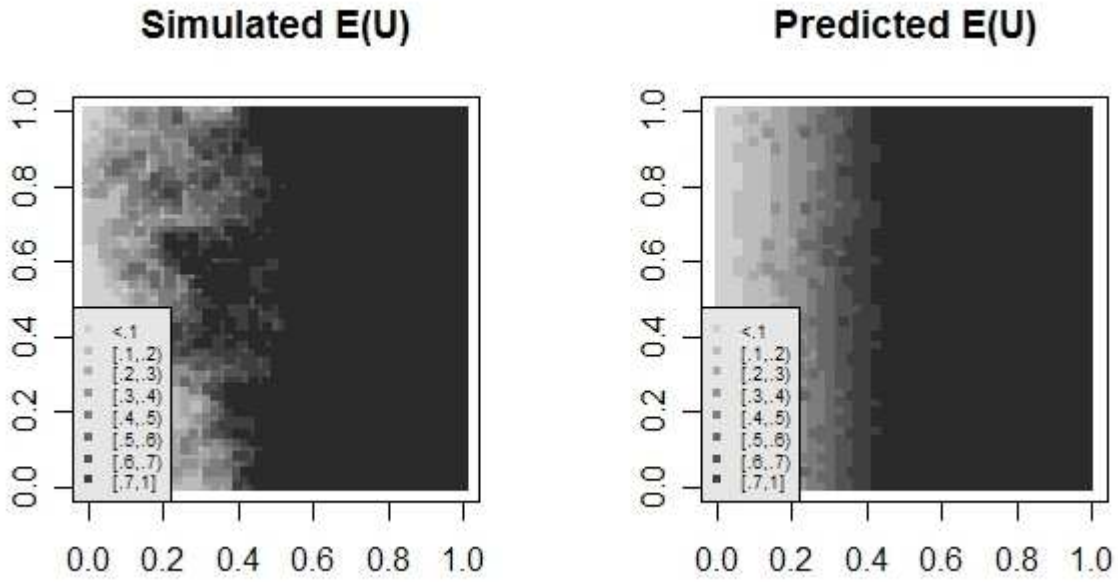

Figure 5: Simulated expectation for incidences (E(U)) and predicted expectation for incidences (predicted E(U)) from TSNC model

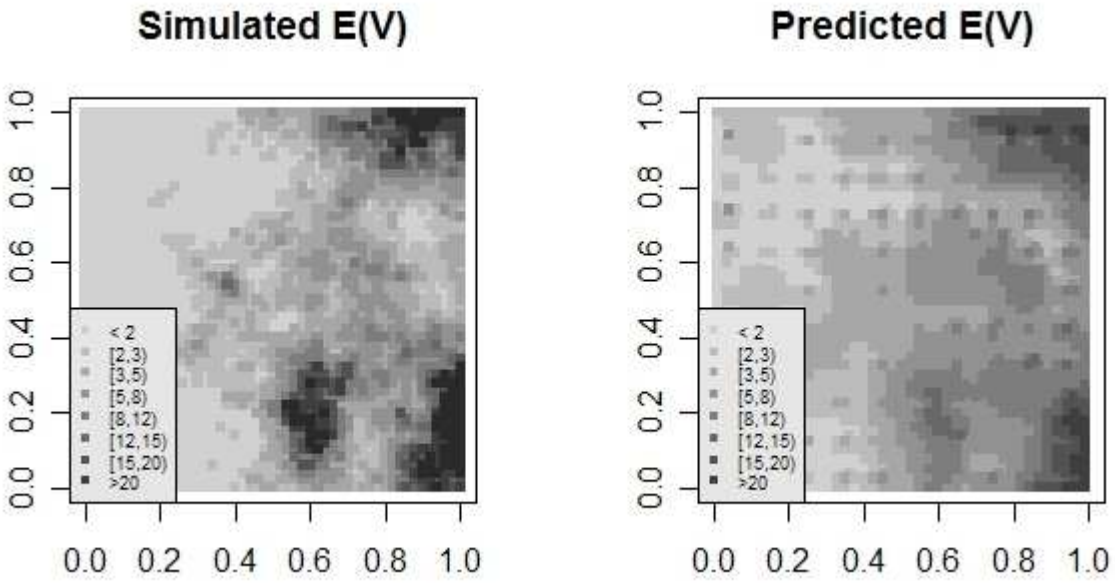

Figure 6: Simulated expected prevalence (E(V)) and predicted expectation for prevalence (predicted E(V)) from TSNC model

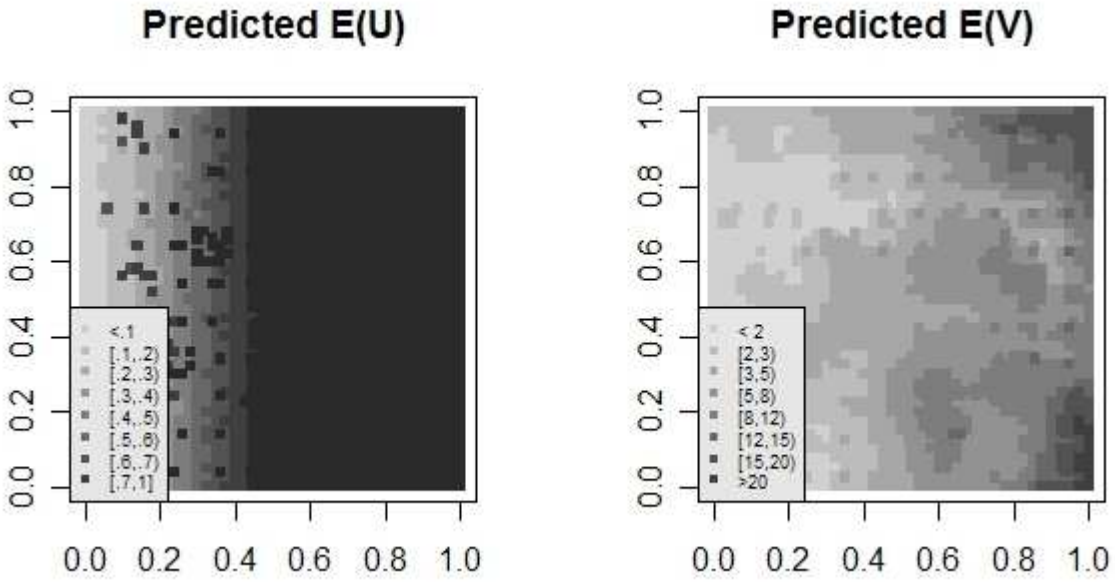

Figure 7: Simulated expectation for incidences (E(U)) and predicted expectation for incidences (predicted E(U)) from from TSIB model

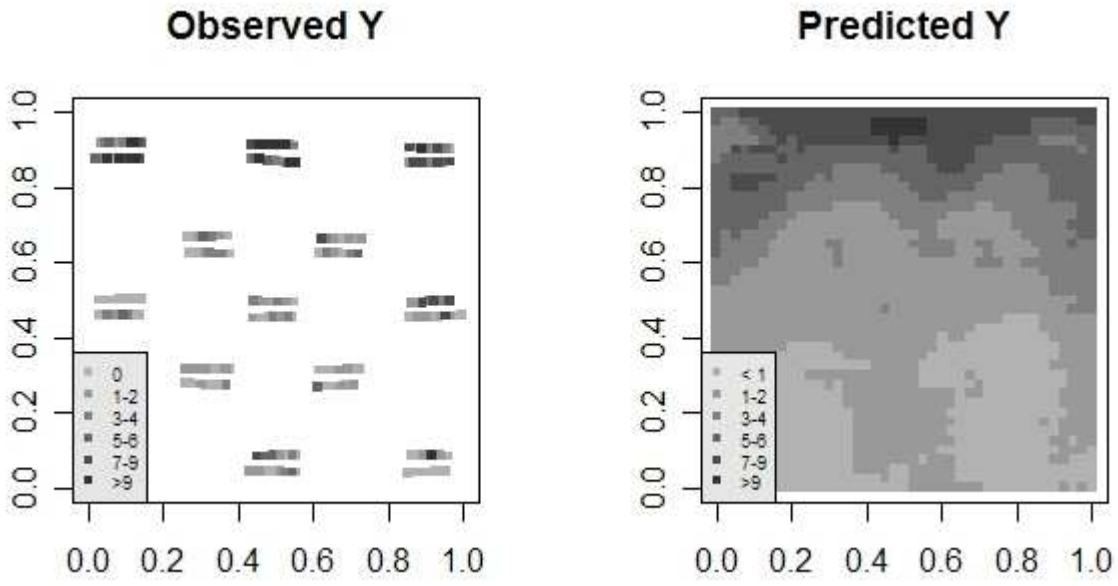

Figure 8: CPB observations (Y) and predicted Y (posterior means) from TSNC model

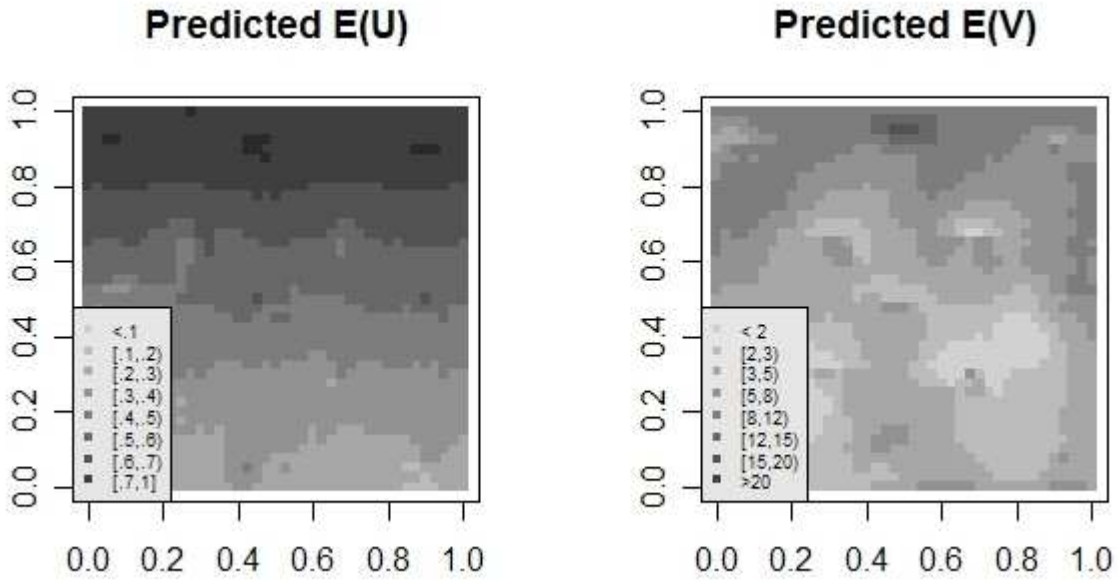

Figure 9: predicted incidence (E(U)) and predicted prevalence (E(V)) (posterior means) for CPB data from TSNC model