The Pennsylvania State University

The Graduate School

Eberly College of Science

A MODEL-BASED ANALYSIS OF SEMICONTINUOUS SPATIAL DATA

A Dissertation in

Statistics

by

Virginia F. Recta

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2009

The dissertation of Virginia F. Recta was reviewed and approved* by the following:

James L. Rosenberger
Professor of Statistics
Dissertation Co-advisor
Co-chair of Committee

Murali Haran
Assistant Professor of Statistics
Dissertation Co-advisor
Co-chair of Committee

Joseph L. Schafer
Associate Professor of Statistics

Shelby Jay Fleischer
Professor of Entomology

Bruce G. Lindsay
Willaman Professor of Statistics
Head of the Department of Statistics

*Signatures are on file in the Graduate School

ABSTRACT

We consider the problem of modeling point-level ('geostatistical') spatial count data with a large number of zeros. We develop a model that is compatible with the scientific assumptions about the data generating process. We use a two-stage spatial generalized linear mixed model framework for the counts, modeling incidence, resulting in 0-1 outcomes, and abundance, resulting in positive counts, as separate but dependent processes, and utilize a bivariate Gaussian process model for characterizing the underlying spatial dependence. We describe a Bayesian approach and study several variants of our two-stage model, consisting of varying covariance and cross-covariance structures for the underlying bivariate Gaussian random process. 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 automated. Finally, we demonstrate the application of our modeling and computational approach on both simulated data and a real data set from an ecological study and compare the performance of the various two-stage models based on inference and prediction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work would not have been possible without the guidance and support of all members of my Thesis Committee. Dr. James Rosenberger was the steady hand through all these years and never gave up on me even when I certainly gave him enough reason to do so. Dr. Murali Haran provided tremendous energy and a clear vision of what this work can add to the study of spatial modeling. Dr. Joseph Shafer made very helpful comments and suggestions that greatly improved clarity and the interpretation of our findings. Dr. Shelby Fleischer provided much insight on the relevance of this work and other applications.

Special thanks to my former professor Dr. Mark Handcock who gave me the idea of using a two-part approach for spatial data with excess zeros and reviewed earlier versions of this work.

Dr. Paul Blom and Dr. Shelby Fleischer were patient and gracious in educating me about the Colorado potato beetle and allowing me to use their data for this study.

Dr. Edward Rajotte was instrumental in starting us on this wonderful adventure, helping me get admitted to this great university and providing much needed support for me and family as we adjusted to our new environment.

Friends and colleagues at the Food and Drug Administration gave constant and unqualified support and encouragement: Anna Nevius, Mary Bartholomew, Veronica Taylor, Laura Stets, Janice Derr, and Joan Buenconsejo.

This whole project was a family undertaking. My husband Jun took on much more than his share of family responsibilities, and my children Miña and JP had to make do with a lot less "Mommy time" for a really... long.... time... to make this happen.

I thank you all from the bottom of my heart.

Chapter 1

Introduction

In many applications it is reasonable to assume that observations distributed across geographical space will likely be correlated. This is very natural in the environmental setting, where spatial structure is induced by various energy inputs (light, thermal, chemical, etc.) resulting in patchy structures (Legendre and Fortin, 1989). In turn, these biotic processes induce aggregation or gradients of organisms and other phenomena. Therefore an ecological phenomenon located at a given sampling point may be similar to other points close by, or even some distance away.

Spatial models account for this dependence by assuming that correlation is present in all directions and that this dependence weakens with increasing distance between data points. Spatial models emerged in part from an area of study known as geostatistics, a collection of methods whose main objective is to predict an unobserved value of an underlying process in continuous space.

1.1 Classical Geostatistics

In classical geostatistical methodology, we assume that the data $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$ is observed at a finite number of sampling locations in two-dimensional space $\{x_{\mathrm{i}}: x \in A, i = 1, \ldots, n\}$, and that the data was generated according to the model

$$Y_i = \mathbf{d}(x_i)^T \boldsymbol{\beta} + S(x_i) + Z_i \qquad \textbf{(1.1)}$$

where $\mathbf{d}(x)$ is a known vector of explanatory variables, $\boldsymbol{\beta}$ is a vector of fixed parameters, $S(x)$ is an unobserved Gaussian process with E[$S(x)$] = 0 and cov{$S(x)$, $S(x')$} = $\sigma^2\rho(x, x')$ and the Z_i are mutually independent N(0, τ^2). Thus, Y- $\mathbf{d}(x)^T\boldsymbol{\beta}$ can be regarded as a 'noisy' version of the underlying spatial stochastic process $S(x)$. Equivalently, we can also state that, conditionally on $S(\cdot)$, the Y$_i$ are mutually independent with the following distribution:

$$Y_i \middle| S(x_i) \overset{ind}{\sim} N\left(\mathbf{d}(x_i)^T \boldsymbol{\beta} + S(x_i),\ \tau^2\right) \qquad \textbf{(1.2)}$$

Predicting $S(x_0)$, the realized value of S at an arbitrary location x_0, is at the center of a procedure called kriging, also called spatial smoothing or spatial interpolation. Under the preceding assumptions, the kriging predictor that minimizes the prediction mean square error $\mathrm{E}\left[\left(\hat{S}(x_0) - S(x_0)\right)^2\right]$ takes the form

$$\hat{S}(x_0) = \sum_{i=1}^{n} w_i(x_0)\left(y_i - \mathbf{d}(x_i)^T \hat{\boldsymbol{\beta}}\right), \qquad \textbf{(1.3)}$$

where the kriging weights $w_i(x_0)$ are derived from the estimated mean and covariance structure of the data. In practice, the parameter $\boldsymbol{\beta}$ of the mean function is first estimated, followed by the parameters of the covariance function using a data-analytic tool called the variogram. The estimated covariance function is then taken as the true covariance function and used to determine the w_i's.

The current practice of classical kriging has several limitations. First, in estimating the prediction variance, no allowance is made for the fact that the parameters of the covariance functions were estimated. Second, the Gaussian assumption is restrictive, given the current wide use of the method for data in situations where normality does not hold or cannot be verified. Third, the method is intended for predicting realizations and their linear combinations, and not for the general case of predicting (possibly non-linear) functionals of the distribution.

1.2 Model-based Geostatistics

Diggle *et al*. (1998) proposed a model-based approach to predicting non-linear functionals of realized values (e.g., the maximum value over a region, or the probability of exceeding a specified threshold) under possibly non-Gaussian realizations. In the same way that McCullagh and Nelder (1989) extended the normal linear model for independent data using generalised linear models, Diggle *et al*. extended the classical geostatistical methodology for spatial data by relaxing the linear assumptions in Eq. **1.1** as follows:

1. $S(x)$ is a Gaussian process with $E[S(x)] = 0$ and $\text{cov}[S(x), S(x')] = \sigma^2\rho(x, x')$
2. Conditionally on S, the random variables Y_i, $i = 1, \ldots, n$ are mutually independent, with distributions $f_i\{y|\ S(x_i)\} = f(y|M_i)$, where $M_i = E[Y_i|\ S(x_i)]$
3. $h(M_i) = \mathbf{d}(x_i)^{\mathrm{T}}\beta + S(x_i)$ for some known link function h, explanatory variables $\mathbf{d}(x_i)$ and parameters $\boldsymbol{\beta}$

Therefore, the unobserved spatial process $S(x)$ is still intrinsically Gaussian, but the observed process Y is no longer linear in S. It is the *expression* of the Gaussian quantity $h(M_i) = \mathbf{d}(x_i)^{\mathrm{T}}\boldsymbol{\beta} + S(x_i)$ that is no longer Gaussian, unless h is the identity link and S has a Gaussian density function. Conditional on the latent process S, the $Y(x_i)$ are independent with distributions $f_i\{y| S(x_i)\}$ in the exponential family. Y is still dependent on S but only through a link function h, where $h(E[Y_i| S(x_i)]) = \mathbf{d}(x_i)^{\mathrm{T}}\boldsymbol{\beta} + S(x_i)$. The authors used Markov chain Monte Carlo (MCMC) techniques to estimate and make inferences about the parameters, predict realizations at arbitrary locations, and estimate non-linear functionals of the posterior distribution.

1.3 An Extension to Semicontinuous Variables

As a further extension to non-Gaussian realizations, we consider the case of spatially distributed semicontinuous variables. A semicontinuous variable has a portion of responses equal to a single value (usually zero) and a continuous, often skewed, distribution of the remaining values. Two-part models for semicontinuous variables have a long history in economics and policy analysis (see for example, Duan *et al.*, 1983, Manning *et al.*, 1987, and Leung and Yu, 1996). In the longitudinal setting, Olsen (1999) and Olsen and Schafer (2001) cite several examples: adolescent substance abuse, dividend income, and expenditures on durable goods and medical care. In modeling semicontinuous longitudinal data, the authors recoded the semicontinuous response, $Y_{ij,}$ into two variables,

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

and **(1.4)**

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

where $j = 1, \ldots, n_i$ indexes the time points for individual $i = 1, \ldots, m$, and g is a monotone increasing function (e.g., log) that will make V_{ij} approximately Gaussian and satisfy the linear model assumptions. They then fitted a two-part 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 allowed a different set of covariates for each part of the model, i.e., a set of covariates for the probability of nonzero response and another set for the mean of nonzero responses. At the same time, a joint distribution for the random coefficients from each part provided a mechanism for relating the two parts of the model.

Modeling semicontinuous variables 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. These models allow the specification of different covariates for each process as well as a mechanism to relate the two parts.

1.4 Motivating 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 (see, for example, Legendre and Fortin 1989, Schotzko and O'Keefe 1989, Schotzko and Smith 1991, Williams *et al*. 1992, Rossi *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 in the study of Colorado potato beetle (CPB) populations (Blom and Fleischer, 2001; Blom *et al*., 2002). The authors described the spatial dynamics of CPB populations in potato fields. In one experiment, counts of CPB large larvae and other life stages per meter-row were observed weekly in sample locations in a field measuring approximately 85 × 85 meters. The smallest resolution available with the GPS technology in use at the time was one meter. One complication encountered in this study is that, due in part to the level of resolution of the observations, a substantial proportion of the observations was zero. Figure **1.1** shows histograms of some of the weekly observations of densityof large larvae per meter, showing the spikes at zero. The expected counts for a Poisson variable with the same mean as the sample is clearly not a good fit for the observed data.

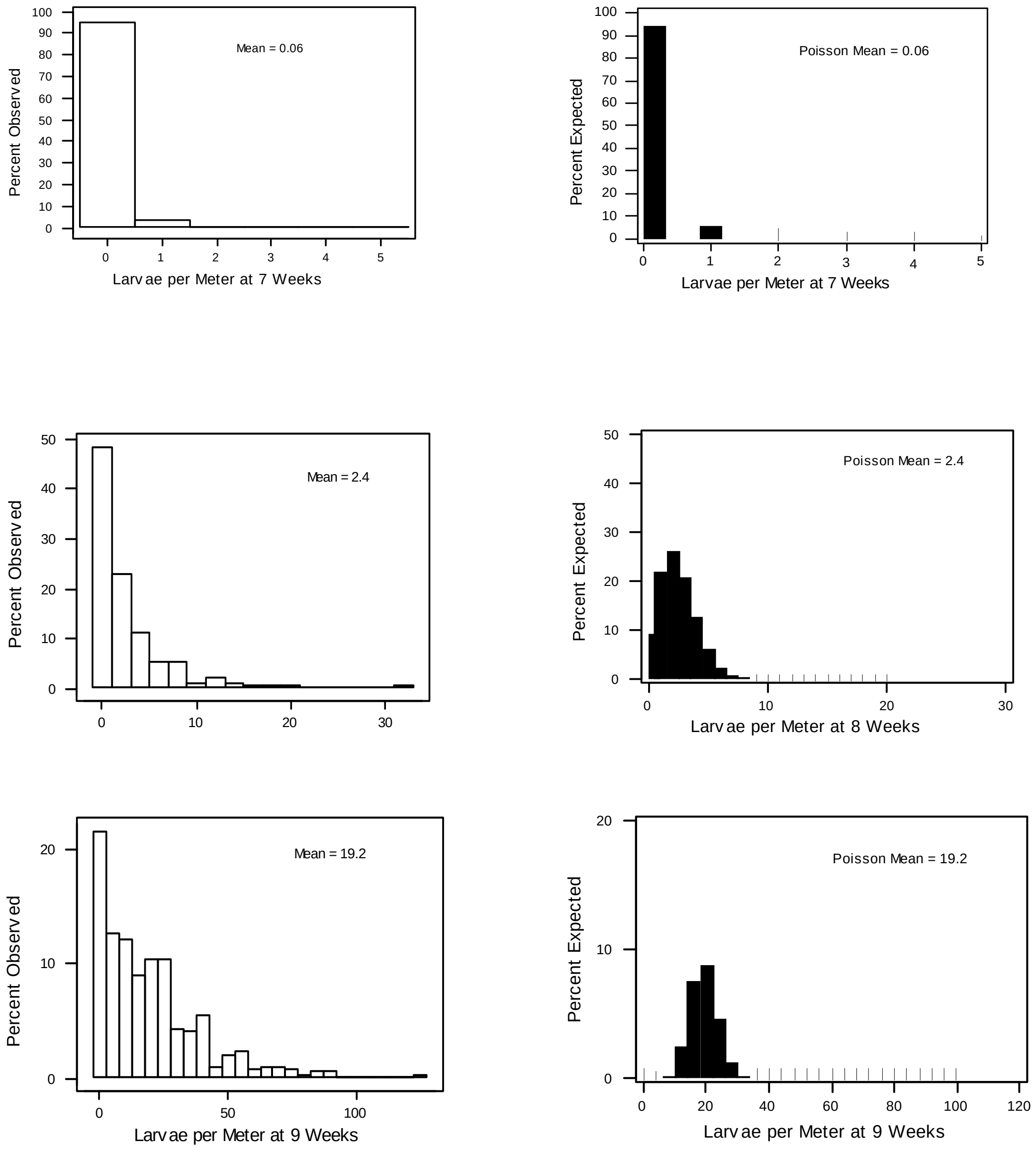


Fig. **1.1**: Number of CPB larvae per meter during weeks 7 to 9, observed (left column) and expected under a Poisson distribution with the same mean (right column).

Some of the other data sets with relatively higher densities appear to have a more standard distribution, possibly Poisson. In these cases, the methods described in Diggle *et al*. (1998) may be suitable. Alternatively, the data may be transformed (e.g., log) into a distribution that is approximately Gaussian and analyzed using trans-Gaussian kriging (Cressie 1993). The distribution may be seen as a manifestation of two biological processes: incidence, as shown by presence or absence; and abundance, as shown by the mean of positive counts. Studying each process separately but simultaneously can be useful, from the point of view of research as well as pest management. At varying times and insect stages, specific interest may also be on various functionals of the distribution, in addition to the usual spatial predictions.

In CPB, there are two life stages of interest: adult and large larvae. CPB is thought to invade new potato crops principally by walking or through short-range flights from nearby sites where they have burrowed over the winter (see, for example, French II *et al*, 1993; Hough-Goldstein and Whalen, 1996; and Lashomb and Ng, 1984). Initially, the whole-field adult density may be low, but considerable interest would be on *where* these adults are, because they are the principal agents for within-field populations to follow. Therefore, the incidence of immigrating adults and the factors affecting it are very important. From the point of view of population studies, it is important to identify the within-field factors that predispose the presence of an immigrating adult. This involves testing hypotheses on the mean part of the process. From the pest management point of view, a relevant functional would be a map of the risk of incidence, which may indicate future population centers. Locations of upper quantiles of severity may also point to nearby origins or sources of overwintered adults (i.e., adults that have spent the winter burrowed in nearby fields, and emerge in the summer to migrate to new crops).

Adults lay eggs, and the eggs hatch and become larvae, the most destructive stage. At this point, incidence and severity are both important, because of the spatial fidelity (i.e., they are usually not mobile) and damage potential of CPB larvae. The spatial dependence between incidence and abundance (e.g., how abundance in one location correlates with incidence in a nearby location) would also be interesting. A map of the mean surface would show areas where the adults appear to have laid eggs. The factors affecting the adults' location choices would be useful from the behavioral point of view. From the management point of view, a map showing the probability of exceeding a mean economic threshold would be needed.

1.5 Objectives

The primary goals of this research are:

1. To develop two-stage models for semicontinuous spatial variables and study the properties of such models; and
2. To develop computationally efficient algorithms to perform inference and prediction for these models.

Chapter 2

Geostatistics Background

2.1 Introduction to Geostatistics and Kriging

Spatial statistics is concerned with the summarization of and inference from data whose spatial dimension is potentially relevant. Spatial statistical methods have been developed under three broad categories (Diggle, 1996): continuous spatial variation, discrete spatial variation, and spatial point processes. Cressie (1993) uses slightly different terminology: geostatistical data, lattice data, and point patterns. In this study, we focus on continuous spatial variation, applicable to any phenomenon in which a random variable of interest, say $Y(x)$, is, in principle, obtainable at any location x within a (typically two-dimensional) spatial region D. The variable Y may itself be continuous, discrete, or categorical, but is presumed to have been generated from a spatially continuous process. Examples of such variables are radiation level that can be measured at any location within a specified region, insect count that can be observed in any location in a field, whether or not a particular organism or species is present in any specific location in a designated area of observation.

In this study we assume an underlying Gaussian process for the random effects in our model. A Gaussian process is an infinite dimensional real valued stochastic process, fully defined by a mean and covariance function, for which every finite dimensional subset has a multivariate normal distribution. Gaussian processes are widely used in geostatistics, where it is common to assume a Gaussian stochastic process *S,* uniquely defined by a mean and covariance function. A single realization produces a surface $S(\cdot)$, and we have finite vector of locations **x** from which we obtain our observations $S(\mathbf{x})$. Cressie (1993) cites two reasons why Gaussian processes are important in geostatistics: "The first is the pragmatic reason that, upon assumption of the Gaussian process, virtually all prediction, estimation, and distribution theory are considerably easier." Using classical statistical methods, spatial analysis for Gaussian random fields is more straightforward. For example, Schabenberger and Gotway (2005) note that the best linear unbiased predictor for $S(x_0)$ at an unobserved location x_0 is generally only best in this restricted class of (linear unbiased) predictors. However, under the Gaussian process assumptions, "this is the best predictor (under squared error loss) among all possible functions of the data." The second reason cited by Cressie (1993) "comes from asymptotic considerations where the net result of many small order (possibly non-Gaussian) effects is approximately Gaussian" (from the Central Limit Theorem).

We introduced classical geostatistics in Section **1.1** and provide more details here. We define $x \in \mathbb{R}^2$ to be a generic data location in two-dimensional Euclidean space and *Y*(*x*) is the observed quantity at location *x*. If $D \subset \mathbb{R}^2$ is a fixed region of interest, and if we let *x* vary continuously over *D*, we generate the multivariate random field or spatial stochastic process $\{Y(x) : x \in D\}$ (Cressie, 1993). We assume that the stochastic process that generated $Y_i = Y(x_i)$ has the form

$$Y_i = \mathbf{d}(x_i)^T \boldsymbol{\beta} + S(x_i) + Z_i \qquad \textbf{(2.1)}$$

with distributional assumptions as described in Section **1.1**. The first set of terms, $\mathbf{d}(x)^T\boldsymbol{\beta}$, represents large-scale variation, also called the *trend*. *S* and *Z* represent the stochastic part of the model. *S* is a zero-mean Gaussian random field with a known covariance function, and is the phenomenon of interest in most applications because it is a key component of prediction. The *Z*'s are independent zero-mean random errors, representing all other factors contributing to variability that do not appear to be spatially relevant. This is also called the *nugget* effect in the literature.

The process $S(x)$ is typically not directly observable. The structure of this latent process can only be inferred from observations on *Y*. Geostatistics is mainly concerned with predicting $S(x_0)$, the value of *S* at an arbitrary location x_0, or some linear functional. Linear predictors for this purpose are called *kriging* predictors. Cressie (1989, 1990) gives a historical account of the development of various forms of kriging.

Diggle *et al*. (1998) summarized the current practice of kriging, and their main points are restated here. Let $\mathbf{Y} = \left(Y(x_1),\ldots,Y(x_n)\right)^T$ be the vector of observations, $\mathbf{M}=\left(\mathbf{d}(x_1)^T\boldsymbol{\beta},\ldots,\mathbf{d}(x_n)^T\boldsymbol{\beta}\right)^T$ be the *n*-element mean vector, **K** be the *n* x *n* covariance matrix with $(i, j)^{\text{th}}$ element $\text{Cov}[S(x_i), S(x_j)]$, and **I** the *n* x *n* identity matrix. Then, by the assumptions stated in Section **1.1**, we have $\mathbf{Y} \sim N\left(\mathbf{M},\ \mathbf{K}+\tau^2\mathbf{I}\right)$. Next, consider $S(x_0)$, the unobserved (i.e., latent) variable at location x_0, and further let **k** be the *n*-element vector with i^{th} element $\text{Cov}[S(x_i), S(x_0)]$, and $\text{k}_0 = \text{Cov}[S(x_0), S(x_0)]$. **Y** and $S(x_0)$ are jointly distributed as

$$\begin{pmatrix} \mathbf{Y} \\ S(x_0) \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mathbf{M} \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{K}+\tau^2\mathbf{I} & \mathbf{k} \\ \mathbf{k}^T & \text{k}_0 \end{pmatrix}\right) \tag{2.2}$$

By well-known properties of the multivariate normal distribution (Anderson, 1958), the conditional distribution of $S(x_0)|\mathbf{Y}$ can be directly deduced from the joint distribution in Equation **2.2**:

$$S(x_0)\,|\,\mathbf{Y} \sim N\left(\mathbf{k}^T\left(\mathbf{K}+\tau^2\mathbf{I}\right)^{-1}(\mathbf{Y}-\mathbf{M}),\ k_0-\mathbf{k}^T\left(\mathbf{K}+\tau^2\mathbf{I}\right)^{-1}\mathbf{k}\right) \qquad \mathbf{(2.3)}$$

The kriging predictor $\hat{S}(x_0)$ is that function of **Y** that minimizes the prediction mean square error $E\left[\left(\hat{S}(x_0)-S(x_0)\right)^2\right]$. That function is $\hat{S}(x_0)=\mathrm{E}\left[S(x_0)\,|\,\mathbf{Y}\right]$, with variance $Var\left[\hat{S}(x_0)\right]=Var\left[S(x_0)\,|\,\mathbf{Y}\right]$, both obtained directly from Equation **2.3**.

In practice, $\boldsymbol{\beta}$ is estimated from the data by assuming that the covariances are known, **K** is nonsingular and $\mathbf{d}(\mathbf{X})$ is of full rank, then computing the generalized least squares (GLS) estimator $\hat{\boldsymbol{\beta}}=\left(\mathbf{d}(\mathbf{X})^T\left(\mathbf{K}+\tau^2\mathbf{I}\right)^{-1}\mathbf{d}(\mathbf{X})\right)^{-1}\mathbf{d}(\mathbf{X})^T\left(\mathbf{K}+\tau^2\mathbf{I}\right)^{-1}\mathbf{Y}$ for use in **M** in Eq. **2.3**. Stein (1999) showed that under these assumptions, the kriging estimator $\hat{S}(x_0)$ with the plug-in $\hat{\boldsymbol{\beta}}$ is the best linear unbiased predictor for $S(x_0)$.

In the kriging predictors above, **K**, **k** and k$_0$ are fixed since the covariance function is assumed known. This is rarely the case in practice, and hence the choice of covariance function and its parameters is pivotal to the kriging process. Typically, the covariance function is chosen by modeling the variogram (or, strictly speaking, the semi-variogram). If the field has a constant mean (i.e., $\mathbf{d}(x_i)^T\boldsymbol{\beta}=\mu$), the variogram quantifies the dependence in the data as a function of distance, *h*, and is defined as

$$\gamma(h)=\frac{1}{2}\ Var\left[Y(x+h)-Y(x)\right] \qquad \mathbf{(2.4)}$$

A method-of-moments variogram estimator due to Matheron (1962) is

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} \left(Y(x_i) - Y(x_j) \right)^2 \qquad \textbf{(2.5)}$$

where $N(h) \equiv \left\{ (i,j) : \| x_i - x_j \| = h \right\}$, the set of all pairs of locations (x_i, x_j) that are a distance *h* (or close to *h*) apart, and |*N*(*h*)| is the number of distinct elements of *N*(*h*). Kitanidis (1997) explains the process of constructing the experimental (or empirical) variogram using Equation **2.5.** Cressie (1993, pages 69-83 and references therein) also describes other methods for variogram estimation.

A parametric family of variograms is chosen based on consistency with the shape of the empirical variogram and what is currently known about the spatial process, such as degree of smoothness and measurement error. The more common forms include linear, spherical, exponential, spherical, Gaussian, or wave, which are described in most standard geostatistics books, for example Cressie (1993), Kitanidis (1997), and Isaaks and Srivastava (1990). The parameters of the chosen variogram are then estimated using various curve fitting methods such as ordinary, generalized and weighted least squares methods as described by Cressie (1993, pages 90-101). The resulting model variogram and its associated covariance function is used to generate **K**, **k** and k_0, and plugged into Equation 2.3 as the known covariance.

However, in the more general case of a nonconstant mean (called *universal kriging*) where both the covariance and regression parameters are unknown, variogram (and covariance) estimation are more involved. One approach (Neuman and Jacobson, 1984) is to estimate $\boldsymbol{\beta}$ using ordinary least squares (i.e., assuming *S*=0), compute a variogram estimator from the residuals, fit a variogram model, then obtain a GLS estimate $\hat{\boldsymbol{\beta}}$ based on the fitted model, and then iterate between updating the variogram and the GLS estimate $\hat{\boldsymbol{\beta}}$. Section 3.4.3 of Cressie (1993) discusses this and other issues in estimating a variogram under universal kriging.

2.2 Alternatives to Classical Geostatistical Methods

Current research in spatial statistics has built upon classical geostatistical methods and has expanded the methodology in several areas. The three areas that concern this study are: (1) selection and parametric estimation of the covariance function as an alternative to the current semi-parametric approach of using variograms; (2) improving estimates of predictive uncertainty by incorporating uncertainty in covariance estimation, and (3) modeling non-Gaussian realizations in space.

2.2.1 Selection and Parametric Estimation of Covariance Functions

As described in Section 2.1, the classical geostatistical practice of choosing the covariance function and estimating its parameters based on variograms proceeds in a rather *ad hoc* manner. However, likelihood-based methods of parameter estimation have also been proposed, most notably maximum likelihood estimation (MLE) and restricted maximum likelihood (REML) estimation.

Vecchia (1988, 1992) propose an iterative procedure for finding MLEs of the covariance parameters to alleviate the large-sample computational limitations of conventional MLE such as those observed by Mardia and Marshall (1984). Mardia and

Watkins (1989) suggest the use of profile log likelihoods as a possible solution to the difficulties pointed out by Warnes and Ripley (1987), such as long ridges in which the likelihood is essentially constant, and convergence to the nearest local maximum when Fisher scoring was used. Likewise, Stein (1999, p. 173) argues that these difficulties are not necessarily shortcomings of the approach; instead, these are potential computational problems when using iterative procedures for finding the maximum when the data provide no information for choosing among the parameter values along the ridge. Like Mardia and Watkins, Stein suggests plotting the log likelihood or judiciously chosen profile log likelihoods to detect these ridges.

Using REML for estimating parameters in spatial covariances was first proposed by Kitanidis (1983). Zimmerman (1989) gives details of REML in observations on a regular lattice and how computations can be reduced in some special cases.

2.2.2 Incorporating Uncertainty in Parameter Estimation

The greater part of studying spatial processes is identifying the covariance structure of the underlying random field *S*, but the fact that this covariance function is almost always estimated is ignored in classical kriging. Estimates of prediction variance do not include uncertainty in the assumed covariance structure, and could be overly optimistic about the precision of the kriging predictors. This is a potential limitation of classical kriging (Diggle *et al.*, 1998).

Bayesian inference provides a way to incorporate parameter uncertainty in prediction by treating the parameters as random variables and integrating over the parameter space to obtain the predictive distribution of any quantity of interest (Ribeiro and Diggle, 1999a). Assuming a Gaussian process, Kitanidis (1986) examined the effect of parameter uncertainty in a Bayesian framework and derived posterior distributions and estimators in the case of (1) known covariance parameters but partially unknown drift coefficients, and (2) unknown drift and partially unknown covariance parameters. Omre (1987) and Omre and Halvorsen (1989) also used Bayesian analysis to incorporate uncertainty in the mean part of the model, and showed that the choice of prior for the mean part defines a continuum of models between simple kriging (constant mean, $\mathbf{d}(x)^T \boldsymbol{\beta} \equiv \mu$) and universal kriging (the linear mean model as defined in Section **1.1**). Cressie (1993, Section 3.4.4) summarized previous work on Bayesian kriging.

Focusing on uncertainty in the covariance parameters, Handcock and Stein (1993) derived Bayes predictive distributions for Gaussian random fields having Matérn covariance functions. The authors showed that when the uncertainty in mean and scale parameters of the covariance function is accounted for, inferences from the resulting Bayesian predictive distribution can differ significantly from those based on the usual plug-in predictive distribution, which raises the question of whether the plug-in predictor is optimal when the covariance function is estimated (as is often the case).

Ribeiro and Diggle (1999a) adopt a fully parametric model-based approach, using hierarchical spatial linear models with specified covariance structures and independent priors for the model parameters. In addition to the usual predictive distributions for values at arbitrary locations, the authors derived posterior distributions for model parameters, under different scenarios of prior knowledge, model choice and degrees of uncertainty.

2.2.3 Modeling non-Gaussian Spatial Data

Up to this point, we have only considered modeling Gaussian realizations of spatially continuous phenomena. The normality assumption conveniently guarantees the optimality of the kriging predictors in Section 2.1 and makes available all the well-known results for Gaussian distributions, so that likelihood-based estimation and Bayesian inference, for instance, are fairly straightforward (Ribeiro and Diggle, 1999a).

For non-Gaussian realizations $\{Y(x): x \in D\}$, and when the objective is to predict the value of $Y(x_0)$, it is fairly common to transform the data so that the resulting variable $Y^*(x) = \phi(Y(x))$ is (approximately) normally distributed, perform kriging in the transformed scale to obtain the optimal predictor $\hat{Y}^*(x_0)$, then back-transform to the original scale to obtain an estimate $\breve{Y}(x_0) = \phi^{-1}\left(\hat{Y}^*(x_0)\right)$. Cressie (1993) showed that this naïve estimate is biased, and, for the lognormal case, described a *lognormal kriging* method that incorporates a bias correction. For a few other transformations, the author refers to Shimizu and Iwase (1987) for closed form expressions of unbiased predictors. *Trans-gaussian kriging*, also described in Cressie (1993), likewise employs standard kriging techniques after applying a normalizing transformation on the non-Gaussian observations, then obtaining an (approximately) unbiased predictor using the δ-method.

Stein (1999), however, observes that a process being “close to” Gaussianity does not guarantee that the linear predictor derived from normality assumptions is a good predictor. Using a Poisson process as an example, he showed that, although in some sense the process is nearly Gaussian, the best linear predictor performs infinitely worse than the best non-linear predictor under the model.

Trans-Gaussian kriging also potentially suffers from the same constraints as classical kriging, primarily underestimation of prediction variance. Additionally, there is an added factor of uncertainty in the choice of the normalizing transformation. De Oliveira *et al.* (1997) extended the work of Handcock and Stein (1993) with the *Bayesian transformed Gaussian* (BTG) approach. In the same way that the Bayesian approach integrated other sources of uncertainty into classical kriging, BTG offers a realistic enhancement to trans-Gaussian kriging, taking into account other major sources of uncertainty, including uncertainty in the choice of a normalizing transformation.

Diggle *et al*. (1998) and Diggle *et al*. (1997) remain within the Bayesian framework but take another approach in dealing with departures from normality. As discussed in Section **1.2**, Diggle *et al.* relaxed the Gaussian assumptions in spatial models by generalizing to realizations in the exponential family, in the same way that McCullagh and Nelder (1989) extended the classical normal linear model to non-normal realizations through generalized linear mixed (GLMM) models. The authors used Markov chain Monte Carlo (MCMC) techniques utilizing Metropolis-Hastings algorithms to estimate and make inferences about the parameters, predict realizations at arbitrary locations, and estimate non-linear functionals of the posterior distribution.

One of the main constraints in the above implementation of spatial GLMM is that the authors used a (Metropolis) fixed-scan algorithm, where the covariance parameters, the regression parameters, and each of the random effects are updated in turn in each scan. This method of updating is computationally intensive because every update of each random effect S_i involves matrix inversions in calculating the conditional variance of this element given the *n-1* other random effects. Building upon Diggle's approach, Christensen *et al*. (2000) and Christensen and Waagepetersen (2002) proposed a more efficient MCMC algorithm using Langevin-Hastings (also called Metropolis-adjusted Langevin) updates. In spatial applications, the Langevin-Hastings algorithm has proved very successful (Møller, 2003) because it simultaneously updates the entire vector of random effects or regression coefficients based on gradient information. The authors used this algorithm and more informative priors to model weed count data.

In the case of spatial logistic modeling for dense point-level binary data Liang *et al.* (2008) proposed modeling at two scales, a macro and a micro scale. In this application, the authors encountered difficulties in estimating spatial correlation parameters and instead used them as tuning constants by fixing them at desired or reasonable values. This work focused more on estimating the regression parameters under a range of fixed correlation parameters.

2.3 Cross-covariance Functions

In our discussion regarding two-part models for semicontinuous variables in Section 1.3 we introduced the notion of recoding the semicontinuous variable Y into two variables (U, V) representing incidence and abundance, respectively, and developing a model for this bivariate response. The proposed two-part spatial model for (U, V) will be described in more detail in Chapter 3, but one of the key components of this model will be a joint specification of the covariance function for the bivariate Gaussian stochastic process (S, Z) where

$$\begin{pmatrix} S \\ Z \end{pmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_S & \mathbf{\Sigma}_{SZ} \\ \mathbf{\Sigma}_{SZ}^T & \mathbf{\Sigma}_Z \end{pmatrix}\right] \quad \textbf{(2.6)}$$

The matrices $\mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_Z$ are the usual covariance matrices in the univariate case, and $\mathbf{\Sigma}_{SZ}$ is the cross-covariance between the two processes *S* and *Z*.

Earth science applications often employ cross-covariance functions in cokriging, also called multivariable spatial prediction (see for example, Webster and Oliver, 2001; Goovaerts, 1998; Stein and Corsten, 1991; Ver Hoef and Barry, 1998; Ver Hoef and Cressie, 1993). A variety of approaches for multivariable spatial data in a hierarchical Bayesian framework are now available. See, for instance, Chapter 7 in Banerjee *et al.* (2004).

A general approach to constructing valid covariance models assumes that the covariance between *S* and *Z* is the sum of several covariance models, so that the covariance function in Eq. **2.6** is constructed as:

$$\begin{bmatrix} \mathbf{\Sigma}_S & \mathbf{\Sigma}_{SZ} \\ \mathbf{\Sigma}_{SZ}^T & \mathbf{\Sigma}_Z \end{bmatrix} = \begin{bmatrix} b_{11}^1 & b_{12}^1 \\ b_{21}^1 & b_{22}^1 \end{bmatrix} g_1\left(\left\|x_i - x_j\right\|\right) + \cdots + \begin{bmatrix} b_{11}^k & b_{12}^k \\ b_{21}^k & b_{22}^k \end{bmatrix} g_k\left(\left\|x_i - x_j\right\|\right), \quad \textbf{(2.7)}$$

where $g_k(\;)$ are various covariance functions and the matrix of coefficients are positive semidefinite. Oliver (2003) considered the difficulties of generating a covariance function using the above formulation, particularly the limitation that they do not allow one to specify different covariance models for the two fields. It is not possible, for example, for one field to have a Gaussian covariance and the other an exponential covariance, unless the two fields are uncorrelated. The author developed an alternative way of constructing valid models for cross-covariance that addresses this limitation.

Given the covariance functions Σ_S and Σ_Z and their correlation $\rho = corr\left(S\left(x_i\right), Z\left(x_i\right)\right)$, a valid cross-covariance model can always be generated by taking $\Sigma_{SZ} = \rho L_S L_Z^T$, where L_S and L_Z are the respective Cholesky factorizations such that $\Sigma_S = L_S L_S^T$ and $\Sigma_Z = L_Z L_Z^T$.

Oliver's approach is quite practical because it allows greater flexibility in the choice of covariance functions while accommodating limited information about the cross-covariance. In many cases the nature of the spatial dependence for 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 this approach these are the only information one needs to construct a valid cross-covariance function: the covariance functions and the correlation between the two variables. We found it useful in deriving the covariance function under the assumption that the *S* random effects are independent, the *Z* random effects have exponential covariance, and the two variables are correlated.

2.4 Models for Semicontinuous Variables

In this section, I examine (1) two classes of models that have been applied to semicontinuous outcomes from cross-sectional studies: two-part models and sample selection models; (2) an extension of the two-part approach to model longitudinal data; and (3) models for spatial data with excessive zeros.

In the econometric literature, two-part models and sample selection models (also called Heckman's model) are widely used in the econometric literature to study semicontinuous outcomes. There is continuing debate on which model is better (see for example, Duan *et al*., 1983, Manning *et al*., 1987, and Leung and Yu, 1996), but the choice largely depends on the parameters of interest as well as the process which generated the semicontinuous data.

Two-part models (Manning *et al*., 1981 and Duan *et al*., 1983) break down the semicontinuous response Y_i, $i = 1, \dots, n$ into two variables (U_i, V_i) where $U_i = I(Y_i > 0)$ and $V_i = Y_i \mid Y_i > 0$. Each variable is then modeled separately, the first a probit equation for U_i and the second a linear model (usually on the log scale) for the positive responses:

$$\begin{aligned} &\Phi^{-1}\left(\Pr(U_i = 1 \mid X_i)\right) = X_i^T \boldsymbol{\alpha} \\ \text{and} \qquad & \qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathbf{(2.8)} \\ &\log\left(V_i \mid X_i\right) = X_i^T \boldsymbol{\beta} + \delta_i \ , \quad \delta_i \sim N\left(0, \sigma^2\right) \end{aligned}$$

where X_i is a vector of explanatory variables, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown parameters. This formulation allows us to study the semicontinuous response Y in two stages: first, the probability of a positive response, then the distribution of Y conditional on it being positive. In addition to its practical appeal, the conditional specification of the variables results in the likelihood being the product of two likelihoods $L_1(\boldsymbol{\alpha})$ and $L_2(\boldsymbol{\beta}, \sigma^2)$, where L_1 is based on the entire sample and L_2 is based on cases having $U_i = 1$. Therefore we can estimate the two models separately because the likelihood is multiplicatively separable, even if the equations in **2.8** are not necessarily independent.

An alternative to the two-equation approach is Heckman's (1974, 1976, 1979) sample selection model. In this specification, semicontinuous responses are due to lack of information on some sector of the population. We are concerned with two processes of interest:

$$\begin{aligned} &Y_i^* = X_i^T \boldsymbol{\beta} + \varepsilon_i \\ \text{and} \qquad & \\ &D_i^* = W_i^T \boldsymbol{\alpha} + \delta_i \end{aligned} \qquad \textbf{(2.9)}$$

where X_i and W_i are vectors of explanatory variables, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown parameters, and ε_i and δ_i are zero-mean error terms with $E[\varepsilon_i \mid \delta_i] \neq 0$. However, Y_i^* and D_i^* are usually latent variables, and the observed sample consists of individuals $i = 1, \ldots, n$ with the following observed variables:

$$\begin{aligned} &Y_i = D_i * Y_i^* \\ \text{and} \qquad & \\ &D_i = \begin{cases} 1 \text{ if } D_i^* > 0 \\ 0 \text{ otherwise} \end{cases} \end{aligned} \qquad \textbf{(2.10)}$$

The value of D_i^* determines whether the variable Y_i^* would be observed $\left(\text{i.e., } Y_i = Y_i^*\right)$, or whether $Y_i = 0$. A common example is the study of female wages (Y^*), where the decision to work might be a function of an unobserved process D^*, and it is clear that some factors would affect both variables. When the value of D^* is such that the female subject decided to work (D =1), then we have a positive outcome and $Y = Y^*$. Otherwise, Y = 0, and we have no information on what non-working female subjects would have earned had they decided to work.

If the process of interest is Y_i^*, the response over the whole population, and all we have is the subsample of Y_i's, Heckman (1974, 1976) showed that OLS estimation of $\boldsymbol{\beta}$ using only the available subsample of positive responses produces biased estimates because of sample selection bias. A number of remedies have been proposed to correct for this bias. Maximum likelihood estimation (Heckman, 1974) relies heavily on the normality of the errors. A two-stage estimation procedure (Heckman, 1979) first models the probability that $D^* > 0$ using the whole sample, then uses the subsample of positive observations and the results in the first stage to consistently estimate the parameters of interest, $\boldsymbol{\beta}$. There are fully parametric and semiparametric versions of these two approaches, reviewed in Vella (1998). Still within the sample selection model, Greene (1994, 1997) developed a parallel method for count data with excess zeros, particularly for Poisson and Negative Binomial regression models.

The sample selection model is suitable when the process of interest is the *unconditional* behavior of the whole population, but we have no information on some individuals because a related process determines whether or not they are observable. Zero outcomes correspond to lack of information on these individuals. In contrast, two-part models are appropriate for studying the process in two stages: first, the process that produces zero vs non-zero outcomes, then the behavior *conditional* on positive outcomes. In two-part models, zeros are real outcomes, not representing insufficient information. Therefore, for the type of spatial phenomena that we are considering in this research, the two-part conditional specification is more suitable than the sample selection approach.

The above argument was Olsen and Schafer's (2001) justification for choosing the two-part approach to model semicontinuous longitudinal data Y_{ij}, where $j = 1, \ldots, n_i$ indexes the time points for individual $i = 1, \ldots, m$. As described in Section 1.3, Y_{ij} was recoded into two variables, U_{ij} and V_{ij}, then fitted with two random effects models, one for the logit probability of $U_{ij} = 1$ and one for the mean conditional response $E(V_{ij}|U_{ij}=1)$:

$$\operatorname{logit}(\pi_i) = X_i \boldsymbol{\alpha} + Z_i c_i \qquad \textbf{(2.11)}$$

$$V_i = X_i^* \boldsymbol{\beta} + Z_i^* d_i + \varepsilon_i \qquad \textbf{(2.12)}$$

where $U_{ij} \sim \text{Bernoulli}(\pi_{ij}),\ \pi_i = (\pi_{i1}, \ldots, \pi_{in_i})^T$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are fixed effects, c_i and d_i are random effects, and $\varepsilon_i \sim N(0, \sigma^2)$. Equation **2.11** is computed on the whole sample, while Equation **2.12** is based only on the observations where $U_{ij} = 1$, so that X_i^* and Z_i^* are the rows of X_i and Z_i that correspond to positive responses in individual i's data matrix. In addition to incorporating random effects to account for individual heterogeneity, the model also allowed for correlation between c_i and d_i:

$$\begin{pmatrix} c_i \\ d_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{cc} & \psi_{cd} \\ \psi_{dc} & \psi_{dd} \end{pmatrix}\right]. \qquad \textbf{(2.13)}$$

Another approach to account for excess zeros in counts is zero-inflated Poisson (ZIP) regression, first proposed by Lambert (1992) in a manufacturing setting. ZIP assumes that a process will be in an imperfect state with probability 1- p_i, during which the distribution of defectives is assumed to be Poisson(λ_i). During this state, zero counts are still possible. However, when the process is in a perfect state with probability p_i, during which it will produce no defectives. The response Y_i is a mixture of 0 with probability p_i and Poisson (λ_i) with probability 1- p_i. The responses $\mathbf{Y} = \left(Y_1,\dots,Y_n\right)^T$ are assumed to be independent with mean vectors $\mathbf{p} = \left(p_1,\dots,p_n\right)^T$ and $\boldsymbol{\lambda} = \left(\lambda_1,\dots\lambda_n\right)^T$ related to the covariates through the canonical link functions $\operatorname{logit}\left(\mathbf{p}\right) = \mathbf{B}\boldsymbol{\beta}$ and $\log\left(\boldsymbol{\lambda}\right) = \mathbf{G}\boldsymbol{\gamma}$, where **B** is the design matrix of covariates for the probability of being in a perfect state and **G** is the design matrix of covariates related to the number of defectives in the imperfect state; these two sets of covariates do not have to be the same. Heilbron (1994) attempted to simplify the two-part model for count data by assuming that the covariates for each part are identical. In zero-altered models, the probability of a zero is a function of the mean of the distribution for the positive counts.

Extensions of the ZIP approach to clustered (Hur *et al.* 2003), longitudinal (Hedeker and Gibbons 2005) and spatial (Agarwal *et al*. 2002) data have been proposed. To model spatial data with excessive zeros, Agarwal *et al*. (2002) formulated a general spatial ZIP model as

$$\log\left(\boldsymbol{\lambda}\right) = \mathbf{B}\boldsymbol{\beta} + \mathbf{W}_1\varphi,\ \ \operatorname{logit}\left(\mathbf{p}\right) = \mathbf{G}\boldsymbol{\alpha} + \mathbf{W}_2\boldsymbol{\gamma}, \qquad \textbf{(2.14)}$$

where **$\boldsymbol{\varphi}$, $\boldsymbol{\gamma}$** are spatial random effects and $\mathbf{W_1}$, $\mathbf{W_2}$ are appropriate incidence matrices. **B** and **G** are specified design matrices which may share common covariates, and $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are their corresponding parameter vectors. The rest of the paper was developed for the case where $\mathbf{W_1{=}I}$ and $\mathbf{W_2{=}0}$, so that the spatial random effects are incorporated only in the positive part of the model. The authors fitted the model within a Bayesian framework. For the spatial random effects, they employed a Markov random field specification and used a conditional autoregressive (CAR) model for the covariance.

Rathbun and Fei (2006) also used the ZIP approach in which the excess zeros are generated by a spatial probit model. Under this model, an excess zero is generated at a given site if the realization of a Gaussian random field falls below a threshold, for instance a minimum measure of habitat suitability in an ecological survey. Using the notation of Eq. **2.14**, the spatial binary field is defined as $\mathbf{Y}= I\left(\mathbf{G}\boldsymbol{\alpha}+\mathbf{W}_2\boldsymbol{\gamma}>\zeta\right)$ where $\boldsymbol{\gamma}$ is a zero-mean Gaussian random field with a specified covariance function. The set of locations where Y_i =1 is interpreted to be the habitat suitable to a species of interest. The spatial ZIP model is then taken as the count Z_i=0 if Y_i= 0, and selecting Z from a Poisson distribution with mean $\lambda_i=\exp\left(B_i'\boldsymbol{\beta}\right)$ if Y_i= 1. The authors fitted the model within a Bayesian framework. The random effects were assumed to vary continuously in space, with a covariance function from the Matérn class.

It is important to note that a ZIP model is a mixture of two distributions: a Poisson (or other common distribution) and another that is degenerate at zero. This model is appropriate when there are two mechanisms that can produce a zero observation. First, zero observations may arise from the distribution degenerate at zero, for instance if a sampled habitat was unsuitable for the species of interest. A zero observation may also arise from the Poisson distribution, for instance if the habitat is suitable but the organism was not present at the time of observation, or was present but not detected because of observer error. In ecological modeling, zero observations from the non-degenerate distribution have been called false negative observations because the organism is present but not counted.

An alternative to the ZIP approach is a two-part model (also called two-stage or hurdle model), which we discussed earlier in this subsection in the context of econometric data. In a hurdle model, one part models the probability of "clearing the hurdle" and generating the non-zero count, and the other part is a zero-truncated distribution for the positive observations. The observed semi-continous variable is still from a mixture model because one model generates zeros and another models the distribution of the positive observations. However, in this approach all zeros are considered true negative observations, i.e., in an ecological setting, the organism is truly not present. These and other approaches to modeling ecological data with excess zeros are discussed in detail in Martin *et al*. (2005), Ridout *et al*. (1998), Potts and Elith (2006), and Tu (2002).

Ver Hoef and Jansen (2007) include space-time random effects to investigate haul-out patterns of harbor seals on glacial ice, and compared the hurdle model to other spatial ZIP models. The authors implemented the models in the context of Gaussian Markov random field models for areal (aggregated) or lattice data. Markov random fields deals with stochastic processes defined on a countable index of spatial sampling units, typically defined by a partitioning of a continuous region into politically or geographically designated sub-regions, such as counties in a state or pixels in an image. The authors fitted the models within a Bayesian framework, using a CAR model for the spatial random effects and a first-order autoregressive model (AR1) to account for temporal dependence.

Chapter 3

The Two-part Spatial Model

3.1 The Model

Consider the semicontinuous response at sample location x_i, $Y_i = Y\left(x_i\right), i = 1, \ldots, n$.

We decompose Y_i into two variables, a binary part and a discrete (or continuous) part:

$$U_i = \begin{cases} 1 & \text{if } Y_i > 0 \\ 0 & \text{if } Y_i = 0 \end{cases}$$

and **(3.1)**

$$V_i = \begin{cases} Y_i & \text{if } Y_i > 0 \\ \text{irrelevant} & \text{if } Y_i = 0. \end{cases}$$

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.

Analogous to the generalized spatial model approach of Diggle *et al*. (1998), we condition on the bivariate spatial stochastic processes $S(x)$ and $Z(x)$. For a finite set of locations $\mathbf{x} = \left(x_1, \ldots, x_n\right)$, these variables are distributed as

$$\begin{pmatrix} S\left(\mathbf{x}\right) \\ Z\left(\mathbf{x}\right) \end{pmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_S & \mathbf{\Sigma}_{SZ} \\ \mathbf{\Sigma}_{SZ}^T & \mathbf{\Sigma}_Z \end{pmatrix}\right] \qquad \textbf{(3.2)}$$

The matrices $\mathbf{\Sigma}_S$ (n x n) and $\mathbf{\Sigma}_Z$ (n_1 x n_1) are the usual covariance matrices in the one-part case, and the cross-covariance matrix $\mathbf{\Sigma}_{SZ}$ (n x n_1) accounts for the relationship between the two processes S and Z.

Conditional on S and Z, U and V are mutually independent, with distributions

$$\begin{aligned} f_{S,i}\left(u \mid S\left(x_i\right)\right) &= f_S\left(u; A_i\right), \text{ where } A_i = \mathrm{E}\left[U_i \mid S\left(x_i\right), \boldsymbol{\alpha}\right] \\ f_{Z,i}\left(v \mid Z\left(x_i\right)\right) &= f_Z\left(v; B_i\right), \text{ where } B_i = \mathrm{E}\left[V_i \mid Z\left(x_i\right), \boldsymbol{\beta}\right]. \end{aligned} \quad \textbf{(3.3)}$$

As in the generalized linear mixed model for independent variables, U and V depend on the unobserved Gaussian spatial process only through their respective expected values A_i and B_i, where

$$\begin{aligned} h_S\left(A_i\right) &= \mathbf{d}_S\left(x_i\right)^T \boldsymbol{\alpha} + S\left(x_i\right) \\ h_Z\left(B_i\right) &= \mathbf{d}_Z\left(x_i\right)^T \boldsymbol{\beta} + Z\left(x_i\right) \end{aligned} \quad \textbf{(3.4)}$$

for some known link functions $h_S(\cdot)$ and $h_Z(\cdot)$, vectors of known explanatory variables $\mathbf{d}_S(\cdot)$ and $\mathbf{d}_Z(\cdot)$ with dimensions p and k, respectively, and vectors of fixed effects $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of dimensions p and k, respectively.

3.2 Preliminary Model for the Motivating Example

For the CPB application discussed in Section 1.4, we propose a Bernoulli distribution for the binary part of the model, truncated Poisson for the discrete part, and an exponential covariance function. Keeping the variable definitions in Equation 3.1,

$$\begin{aligned} U_i &\sim \text{Bernoulli}\left(A_i\right),\ A_i = \Pr\left[U_i = 1 \mid S\left(x_i\right), \boldsymbol{\alpha}\right] \\ V_i &\sim \text{ TruncPoisson}\left(B_i\right),\ \frac{B_i}{1-e^{-B_i}} = \mathrm{E}\left[V_i \mid Z\left(x_i\right), \boldsymbol{\beta}\right] \end{aligned} \quad \textbf{(3.5)}$$

and subsequently link these realizations to the underlying Gaussian process through

$$\begin{aligned} \operatorname{logit}(A_i) &= \mathbf{d}_S(x_i)^T \boldsymbol{\alpha} + S(x_i) \\ \log(B_i) &= \mathbf{d}_Z(x_i)^T \boldsymbol{\beta} + Z(x_i) \end{aligned} \qquad \textbf{(3.6)}$$

We note that a truncated Poisson variable has probability distribution

$$\begin{aligned} P(V_i = v \mid B_i) &= \frac{\exp(-B_i)B_i^v}{(1-\exp(-B_i))v!} \\ &= \exp\left[v\log(B_i) - \log(\exp(B_i)-1) + \log(v!)\right] \end{aligned} \qquad \textbf{(3.7)}$$

which is in canonical form, with the logarithmic function as canonical link.

Finally, we assume exponential covariance for *S* and *Z*,

$$\begin{aligned} \operatorname{cov}(S(x_i), S(x_j)) &= \sigma_{\mathrm{S}}^2 \exp\left[-\theta_{\mathrm{S}} \left\|x_i - x_j\right\|\right] \\ \operatorname{cov}(Z(x_i), Z(x_j)) &= \sigma_{\mathrm{Z}}^2 \exp\left[-\theta_{\mathrm{Z}} \left\|x_i - x_j\right\|\right] \end{aligned} \qquad \textbf{(3.8)}$$

for some $\sigma_S^2 > 0,\ \sigma_Z^2 > 0,\ \theta_S > 0$ and $\theta_Z > 0.$ The cross-covariance function is constructed as described by Oliver (2003) by taking $\Sigma_{SZ} = \rho_{SZ} L_S L_Z{}^T$ where L_S and L_Z are the respective Cholesky factors where $\Sigma_S = L_S L_S^T$, and $\Sigma_Z = L_Z L_Z^T$, and $\boldsymbol{\rho}_{\mathrm{SZ}}$ is the correlation between S and Z at the same location $\left(\text{i.e., } \rho(S(x_i), Z(x_i))\right)$. In this formulation, therefore, there are three unknown parameter vectors: $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{p-1})^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{k-1})^T$, and $\boldsymbol{\gamma} = (\theta_S, \theta_Z, \rho_{SZ}, \sigma_S^2, \sigma_Z^2)$.

3.3 Model Features

The model as described has several desirable features, due mainly to its flexibility. A two-part model allows us to examine the features of each component of a semicontinuous 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 (i.e., 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.

The underlying spatial process is still Gaussian, so this is not a major departure from the present practice of classical geostatistics. However, 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 $\boldsymbol{\Sigma}_{SZ}$ 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, $U(x_j)$.

Our approach differs from the ZIP models employed by recent works of Agarwal *et al*. (2002) and Rathbun and Fei (2006) with respect to model construction as well as incorporation of spatial dependence. As we describe in Section 2.4, the zero observations in ZIP models can come from true or false negatives, whereas a two-part model only has true zero observations. We consider the zero observations in the CPB large larvae count in our application to be true negatives because the larvae are not mobile and are easily detected on potato leaves, and therefore it is highly unlikely that they would be undetected or not present at the time of observation. In this data set we consider the "hurdle" that determines presence or absence to be a combination of habitat suitability as well as whether or not the adult CPB laid eggs on the plant in a specific location..

Our two-stage approach is similar to Ver Hoef and Jansen (2007), separating the binary and count processes. They implement the two-stage model in the context of Gaussian Markov random field models for areal (aggregated) data. Their model is specified in terms of conditional, rather than joint, distributions, incorporating local dependence between spatial units.

Our model differs from Ver Hoef and Jansen (2007) because 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. The geostatistical setting allows us to interpolate realizations in unobserved locations 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. This model is useful in many ecological and biological settings where such data are common. The model is specified jointly, allowing varying degrees of association that is typically a function of distance between pairs of observations. We also include a mechanism to relate the two parts of the model via a cross-covariance between the spatial random effects. However, in our applications in Chapter 5 and Chapter 6, we demonstrate the flexibility of our approach by specifying simpler covariance structures for the random effect (*S, Z*), including two sub-models where *S* and *Z* are independent random vectors, which is the assumption in their model.   In this sense we can consider the Ver Hoef and Jansen (2007) model as a special case of our two-stage model specification.

Chapter 4

Modeling and Computation

4.1 The Distribution of the Semicontinuous Variable

We have two observed vectors $\mathbf{u} = (u_1, \ldots, u_n)^T$ and $\mathbf{v} = (v_1, \ldots, v_{n_1})^T$ whose mean values depend on the regression parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as well as the unobserved and related Gaussian processes $\mathbf{s} = (s_1, \ldots, s_n)^T$ and $\mathbf{z} = (z_1, \ldots, z_n)^T$. Under the assumptions given in Section **3.1**, the unconditional density of (**u**, **v**) is given by the (*n*+*n₁*)-fold integral

$$f(\mathbf{u}, \mathbf{v}) = \int \left[\prod_{i=1}^{n} f_S(u_i \mid \boldsymbol{\alpha}, s_i) \prod_{i=1}^{n_1} f_Z(v_i \mid \boldsymbol{\beta}, z_i) \right] g_{n,n_1}(\mathbf{s}, \mathbf{z}) d\mathbf{s}\, d\mathbf{z} \qquad \textbf{(4.1)}$$

where $d\mathbf{s} = \prod_{i=1}^{n} ds_i, d\mathbf{z} = \prod_{i=1}^{n_1} dz_i$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{p-1})^T$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{k-1})^T$. The function $g_{m,m_1}(s, z)$ is the (*m*+*m₁*)-dimensional multivariate normal probability density of the first *m* and *m₁* elements of **s** and **z**, respectively. For instance, in our proposed model for the motivating example in Section **3.2**, (**S**, **Z**) are normally distributed with zero mean and covariance matrix determined by the parameter vector $\boldsymbol{\gamma} = (\theta_S, \theta_Z, \rho_{SZ}, \sigma_S^2, \sigma_Z^2)$. In Equation **4.1** and throughout this text we also implicitly condition on the covariates $\mathbf{d}_S(x)$ and $\mathbf{d}_Z(x)$, which are assumed known.

Given the underlying Gaussian variables (**S**, **Z**), the (U_i, V_i) are independent. Moreover, because of the conditional specification of the two-part model, their likelihoods given (**S**, **Z**) are multiplicatively separable.

The unconditional density in Equation **4.1** will be our basis for estimating parameters and predicting values at arbitrary locations. Some work on this is presented in the succeeding sections.

4.2 Choice of Priors

In the case of spatial GLMM applications, there is limited guidance on choosing priors, and even less so in our case of a two-part spatial GLMM.

We used independent priors for all parameters. For the covariance parameters, we use log-uniform proper priors for $\left(\theta_S,\ \theta_Z\right)$ and uniform proper priors for $\left(\rho_{SZ}, \sigma_S, \sigma_Z\right)$. The log-uniform prior on a finite interval, $\pi\left(\theta\right) \propto \theta^{-1}, \log\theta \in \left[t_1, t_2\right]$ was employed by Christensen *et al*. (2000) and they show that, along with other conditions, the posterior is proper in the case of the spatial GLMM for Poisson observations. Stein (1998) points out that one reason for the fat upper tail of the posterior distribution of θ in the exponential covariance function $\operatorname{cov}(S(x_i), S(x_j)) = \sigma_S^2 \exp\left[-\theta_S \left\|x_i - x_j\right\|\right]$ under the uniform prior used by Diggle *et al*. (1998) may be that once θ exceeds a certain value, the observations are essentially uncorrelated, so that further increases in θ have almost no effect on the correlation matrix. We avoid this by choosing a prior that reduces the likelihood of larger values of θ. For $\left(\sigma_S, \sigma_Z\right)$, uniform priors on the square root of variance parameters in hierarchical models has recently been suggested by Gelman (2006). Likewise, Ver Hoef and Jansen (2007) use diffuse uniform priors for all σ, to keep random effects from becoming too large and causing numerical instability. They also put a uniform prior [0, 1] on all autoregression parameters.

We used improper uniform priors for all regression parameters $\boldsymbol{\alpha} = \left(\alpha_0, \alpha_1, \ldots, \alpha_{p-1}\right)^T$ and $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_{k-1}\right)^T$ to reflect lack of prior information.

4.3 Prediction, Estimation and Inference

Suppose we wanted to predict S and Z at an arbitrary location x_0. The unconditional density of (S_0, Z_0, **U**, **V**) is given by the (n+n_1)-fold integral

$$f(s_0, z_0, \mathbf{u}, \mathbf{v}) = \int \left[\prod_{i=1}^{n} f_S(u_i \mid \boldsymbol{\alpha}, s_i) \prod_{i=1}^{n_1} f_Z(v_i \mid \boldsymbol{\beta}, z_i) \right] g_{n+1, n_1+1}(\mathbf{s}, s_0, \mathbf{z}, z_0) d\mathbf{s}\, d\mathbf{z} \qquad \textbf{(4.2)}$$

Therefore, the conditional density of $\left(S(x_0), Z(x_0) \mid \mathbf{U}, \mathbf{V}\right)$ is the ratio of Equations **4.1** and **4.2,**

$$f\left(s_0, z_0 \mid \mathbf{u}, \mathbf{v}\right) = \frac{f\left(s_0, z_0, \mathbf{u}, \mathbf{v}\right)}{f\left(\mathbf{u}, \mathbf{v}\right)}. \qquad \textbf{(4.3)}$$

At this arbitrary location, the generalized linear predictor and prediction variance for S_0 are

$$\begin{aligned} \hat{S}(x_0) &= E\left(S_0 \mid \mathbf{u}, \mathbf{v}\right) \\ &= \frac{\int\int s_0 f\left(s_0, z_0, \mathbf{u}, \mathbf{v}\right) ds_0 dz_0}{f\left(\mathbf{u}, \mathbf{v}\right)} \end{aligned} \qquad \textbf{(4.4)}$$

$$Var\left(\hat{S}(x_0)\right) = E\left(S_0^2 \mid \mathbf{u},\mathbf{v}\right) - \left[E\left(S_0 \mid \mathbf{u},\mathbf{v}\right)\right]^2$$

$$= \frac{\int\int s_0^2 f\left(s_0, z_0, \mathbf{u}, \mathbf{v}\right) ds_0 dz_0}{f\left(\mathbf{u},\mathbf{v}\right)} - \left[\frac{\int\int s_0 f\left(s_0, z_0, \mathbf{u}, \mathbf{v}\right) ds_0 dz_0}{f\left(\mathbf{u},\mathbf{v}\right)}\right]^2, \qquad \textbf{(4.5)}$$

and similarly for Z_0. It is clear that since the distribution of (**U**, **V**) is a complex multivariate function, expressions Equations 4.4 and 4.5 are analytically intractable, except when we use the identity link.

In addition to predicting at arbitrary locations, we are also interested in estimating the regression parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and testing the significance of individual coefficients. It is important to note that these parameters have a conditional interpretation, in that $\boldsymbol{\alpha}$ reflects the effect of the covariates $\mathbf{d}_S\left(x_i\right)$ on $E\left[U_i \mid S_i\right]$ and $\boldsymbol{\beta}$ the effect of covariates $\mathbf{d}_Z\left(x_i\right)$ on $E\left[V_i \mid Z_i, U_i = 1\right]$.

Properties of **S** and **Z**, as individual and joint processes, are also of interest. Separately, the covariance functions of **S** and **Z** reveal the spatial persistence of incidence and abundance, respectively. Jointly, we can study the cross-covariance structure of these two processes. Finally, other quantities of interest would be functionals of the distribution itself. 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.

In all the above quantities of interest, the complexity of the distribution rules out closed form expressions for estimators and standard errors.

4.4 Computational Strategies

4.4.1 General Approach

We can decompose the model into its different components and dependencies, as in Figure 4.1. Recall that $\boldsymbol{\gamma}$ is the vector of parameters associated with the selected covariance function, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the regression parameters for the means of **U** and **V**, respectively. **U** depends on $\boldsymbol{\alpha}$ and the unobserved vector **S**, and likewise **V** depends on $\boldsymbol{\beta}$ and **Z**. The node (S_0, Z_0) is relevant only when the interest is on prediction. Notice that $\boldsymbol{\gamma}$ is conditionally independent of **U** and **V** given (**S**, **Z**), and therefore all the spatial dependence is contained in the relationship between **S** and **Z**, which are normally distributed. This specification allows us to model spatial dependence in the conventional geostatistical framework.

Models of this kind are in a class known as generalized linear mixed models or GLMM (Breslow and Clayton, 1993), generalized linear models that include one or more random effects. A popular approach is to use a Bayesian version of the GLMM, by specifying priors for the parameters. Markov chain Monte Carlo (MCMC) algorithms can then be used to fit such Bayesian models (Zeger and Karim, 1991; Clayton, 1996; Olsen, 1999). For spatial data, Clayton and Kaldor (1987) and Besag *et al.* (1991) mapped disease risk, incorporating spatial dependence among the discrete spatial regions by assuming a Markov random field model and then calculating the posterior distributions of quantities of interest using Gibbs sampling (Geman and Geman, 1984).

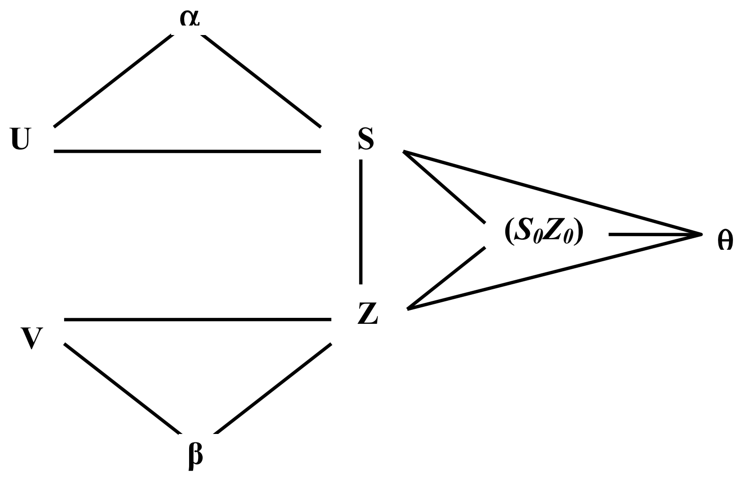


Fig. **4.1**: The components of the model and their structural dependencies. (Adapted from Diggle *et al*., 1998)

Diggle *et al.* (1998) and Christensen and Waagepetersen (2002) used Bayesian inference via MCMC to apply generalized linear mixed models (GLMM) to geostatistical data, incorporating spatial dependence by assuming that the random part of the model is a spatially continuous Gaussian process. The use of MCMC for estimation and prediction in model-based geostatistics, specifically with GLMM for spatial data, is presented well in Diggle *et al.* (2002).

To fit the two-part spatial GLMM model, we also took a Bayesian approach to estimation and inference. We used MCMC techniques to simulate realizations from our complex posterior distribution by repeatedly sampling from the more tractable conditional and marginal distributions.

From a Bayesian perspective, we are interested in the joint posterior distribution $P\left(\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{S},\mathbf{Z}\mid\mathbf{U},\mathbf{V}\right)$ for inference and $P\left(S_0,Z_0\mid\mathbf{U},\mathbf{V},\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{S},\mathbf{Z}\right)$ for prediction. Assuming independent priors, the posterior distributions are fully characterized by these conditional distributions:

$$\begin{aligned} P\left(\mathbf{U}\mid\mathbf{V},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{S},\mathbf{Z},\boldsymbol{\gamma}\right) &= P\left(\mathbf{U}\mid\boldsymbol{\alpha},\mathbf{S}\right) \\ &= \prod_{i=1}^{n} f_S\left(u_i\mid\boldsymbol{\alpha},S_i\right) \end{aligned} \qquad \textbf{(4.6)}$$

$$\begin{aligned} P\left(\mathbf{V}\mid\mathbf{U},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{S},\mathbf{Z},\boldsymbol{\gamma}\right) &= P\left(\mathbf{V}\mid\boldsymbol{\beta},\mathbf{Z}\right) \\ &= \prod_{i=1}^{n_1} f_Z\left(v_i\mid\boldsymbol{\beta},Z_i\right) \end{aligned} \qquad \textbf{(4.7)}$$

$$\begin{aligned} P\left(\boldsymbol{\gamma}\mid\mathbf{U},\mathbf{V},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{S},\mathbf{Z},\right) &= P\left(\boldsymbol{\gamma}\mid\mathbf{S},\mathbf{Z}\right) \\ &\propto P\left(\mathbf{S},\mathbf{Z}\mid\boldsymbol{\gamma}\right)P\left(\boldsymbol{\gamma}\right) \end{aligned} \qquad \textbf{(4.8)}$$

$$\begin{aligned} P\left(\boldsymbol{\alpha}\mid\mathbf{U},\mathbf{V},\boldsymbol{\beta},\mathbf{S},\mathbf{Z},\boldsymbol{\gamma}\right) &= P\left(\boldsymbol{\alpha}\mid\mathbf{U},\mathbf{S}\right) \\ &\propto P\left(\mathbf{U}\mid\boldsymbol{\alpha},\mathbf{S}\right)P\left(\boldsymbol{\alpha}\right) \end{aligned} \qquad \textbf{(4.9)}$$

$$\begin{aligned} P\left(\boldsymbol{\beta}\mid\mathbf{U},\mathbf{V},\boldsymbol{\alpha},\mathbf{S},\mathbf{Z},\boldsymbol{\theta}\right) &= P\left(\boldsymbol{\beta}\mid\mathbf{V},\mathbf{Z}\right) \\ &\propto P\left(\mathbf{V}\mid\boldsymbol{\beta},\mathbf{Z}\right)P\left(\boldsymbol{\beta}\right) \end{aligned} \qquad \textbf{(4.10)}$$

$$\begin{aligned} P\left(\mathbf{S}\mid\mathbf{U},\mathbf{V},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{Z},\boldsymbol{\gamma}\right) &= P\left(\mathbf{S}\mid\mathbf{U},\boldsymbol{\alpha},\mathbf{Z},\boldsymbol{\gamma}\right) \\ &= P\left(\mathbf{U}\mid\boldsymbol{\alpha},\mathbf{S}\right)P\left(\mathbf{S}\mid\mathbf{Z},\boldsymbol{\gamma}\right) \end{aligned} \qquad \textbf{(4.11)}$$

$$\begin{aligned} P(\mathbf{Z} \mid \mathbf{U},\mathbf{V},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{S},\boldsymbol{\gamma}) &= P(\mathbf{Z} \mid \mathbf{V},\boldsymbol{\beta},\mathbf{S},\boldsymbol{\gamma}) \\ &= P(\mathbf{V} \mid \boldsymbol{\beta},\mathbf{Z})\,P(\mathbf{Z} \mid \mathbf{S},\boldsymbol{\gamma}) \end{aligned} \qquad \mathbf{(4.12)}$$

MCMC enables simulation of outcomes from a desired joint posterior distribution by sampling repeatedly from the conditional and marginal distributions that completely determine the posterior distribution. For our problem we set up a Markov chain with transition probabilities following the conditional distributions in equations **4.8** through **4.12**, then sampled repeatedly from each conditional distribution given the most recent values of the other unknowns, for instance sampling $\boldsymbol{\gamma}^{(k+1)}$ from $P\left(\boldsymbol{\gamma} \mid \mathbf{U}^{(k)},\mathbf{V}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\beta}^{(k)},\mathbf{S}^{(k)},\mathbf{Z}^{(k)}\right)$. After a sufficiently long burn-in period of B iterations, we consider the process to have converged to the desired joint posterior distribution $P(\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{S},\mathbf{Z} \mid \mathbf{U},\mathbf{V})$. This distribution can be approximated by the empirical distribution of M draws, $\left\{\left(\boldsymbol{\gamma}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\beta}^{(k)},\mathbf{S}^{(k)},\mathbf{Z}^{(k)}\right): k = B+r, B+2r, \ldots, B+Mr\right\}$ where r is large enough for the draws to be nearly uncorrelated, and M is chosen to give sufficient precision to the empirical distribution of interest.

4.4.2 The Metropolis-Hastings Sampling Algorithm

To sample from the conditional distributions in equations **4.8** through **4.12**, we used algorithms based on the Metropolis-Hastings (MH) algorithm (Metropolis *et al.*, 1953; Hastings, 1970). MH algorithms generate Markov chains which converge to a target distribution $f(y)$ by successively sampling from an arbitrary proposal distribution $q(y|y^*)$ and imposing a random rejection step at each transition. To simulate samples $y^{(1)},\dots, y^{(k)}$ from $f(y)$, MH entails first simulating a candidate value y^C from $q(y|y^{(j)})$. Next, it takes $y^{(j+1)}=y^C$ with probability

$$\alpha\left(y^{(j)},y^{C}\right)=\min\left\{1,\frac{f\left(y^{C}\right)\,q\left(y^{(j)}\mid y^{C}\right)}{f\left(y^{(j)}\right)\,q\left(y^{C}\mid y^{(j)}\right)}\right\} \qquad \textbf{(4.13)}$$

and set $y^{(j+1)}=y^{(j)}$ otherwise. Here we note that $f(y)$ need only be known up to a constant, which makes the algorithm convenient when sampling from non-standard target distributions such as what we have here. Implementation of MH and other issues related to MCMC are well-developed in the literature (for example, Gilks *et al*., 1996; Chib and Greenberg, 1995; Casella and George, 1992; and Robert and Casella, 1999).

An important decision in implementing the MH algorithm is choosing a candidate-generating transition density $q(y|y^{(j)})$. We used both the Gaussian random walk Metropolis and Langevin-Hastings algorithms, as described below.

4.4.3 Gaussian Random Walk Metropolis

We updated the regression and covariance parameters $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ element-wise using Gaussian random walk Metropolis. In this algorithm, the candidate generating density $q(y|y^{(j)})$ is a normal distribution with mean equal to the current state, $y^{(j)}$ and a user-specified covariance h>0, so that $y^{C}\sim N\left(y^{(j)},h\right)$. An equivalent formulation is to generate a random increment r, where $P(r)=N\left(0,h\right)$ and take $y^{C}=y^{(j)}+w$, so that the process moves (walks) from $y^{(j)}$ by a random distance *w*. Note that because the distribution of *w* is symmetric $\left(P(w)=P(-w)\right)$, then $q\left(y^{(j)} \mid y^{C}\right)=q\left(y^{C} \mid y^{(j)}\right)$ and the acceptance probability in Eq. **4.13** simplifies to

$$\alpha\left(y^{(j)}, y^{C}\right)=\min\left\{1, \frac{f\left(y^{C}\right)}{f\left(y^{(j)}\right)}\right\}. \tag{4.14}$$

For example, to update θ_S at the j^th iteration, we generate *w* as described above then compute $\theta_S^C=\theta_S^{(j)}+w$. We accept this candidate value with probability

$$\begin{aligned}\alpha\left(\theta_S^{(j)}, \theta_S^{C}\right)&=\min\left\{1, \frac{f\left(\theta_S^{C}\right)}{f\left(\theta_S^{(j)}\right)}\right\}\\ &=\min\left\{1, \frac{P\left(\mathbf{S},\mathbf{Z} \mid \theta_S^{C}, \theta_Z^{(j)}, \rho_{SZ}^{(j)}\right)P\left(\theta_S^{C}\right)}{P\left(\mathbf{S},\mathbf{Z} \mid \theta_S^{(j)}, \theta_Z^{(j)}, \rho_{SZ}^{(j)}\right)P\left(\theta_S^{(j)}\right)}\right\}\end{aligned}$$

The updates for all elements of $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ follow similarly, tuning *h* for each element to allow acceptance rates in the range of 0.25-0.60.

4.4.4 The Langevin-Hastings Algorithm

For updating the spatial random effects S and Z, we employed the Langevin-Hastings (LH) algorithm as implemented in Christensen *et al.* (2000) and Christensen and Waagepetersen (2002). In this method, rather than be centered at the current state, the proposal center is adjusted according to the information about where the target density is likely to be greater.

Let **Σ** be the covariance of (**S, Z**) as defined in Eq. 2.2, and let $\mathbf{\Sigma}^{1/2}$ be the square root so that $\mathbf{\Sigma} = \mathbf{\Sigma}^{1/2}\left(\mathbf{\Sigma}^{1/2}\right)^T$. We take (**S, Z**) $= \mathbf{\Sigma}^{1/2}\mathbf{\Gamma}$, where **Γ** follows a (n + n_1)-dimensional standard multivariate Gaussian distribution. Recall that we have the two-part observed vector (**U**, **V**), where $U_i \sim$ Bernoulli(A_i) and $V_i \sim$ TruncPoisson(B_i), with conditional means $A_i = \Pr[U_i = 1 \mid S(x_i), \boldsymbol{\alpha}]$ and $B_i^* = \dfrac{B_i}{1-e^{-B_i}} = \mathrm{E}[V_i \mid Z(x_i), \boldsymbol{\beta}]$. We employ canonical link functions logit(A_i) = $\mathbf{d}_S(x_i)^T \boldsymbol{\alpha} + S(x_i)$ and log(B_i) = $\mathbf{d}_Z(x_i)^T \boldsymbol{\beta} + Z(x_i)$. Let

$$\nabla(\gamma) = \frac{\partial}{\partial\gamma}\log f(\gamma \mid y) = -\gamma + \left(\Sigma^{1/2}\right)^T \left\{ \begin{array}{c} \left\{ \left(u_i - A_i\right) \dfrac{h'_c\left(A_i\right)}{h'\left(A_i\right)} \right\}_{i=1}^{n} \\ \left\{ \left(v_i - B_i^*\right) \dfrac{g'_c\left(B_i\right)}{g'\left(B_i\right)} \right\}_{i=n+1}^{n+n_1} \end{array} \right\} \qquad \mathbf{(4.15)}$$

denote the gradient of the log target density, 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. Since we used canonical links in both cases, $\dfrac{h'_c\left(A_i\right)}{h'\left(A_i\right)}$ = $\dfrac{g'_c\left(B_i\right)}{g'\left(B_i\right)}$ = 1 and the gradient simplifies to

$$\nabla(\gamma) = \frac{\partial}{\partial\gamma}\log f(\gamma \mid y) = -\gamma + \left(\Sigma^{1/2}\right)^T \left\{ \begin{array}{c} \left\{\left(u_i - A_i\right)\right\}_{i=1}^{n} \\ \left\{\left(v_i - B_i^*\right)\right\}_{i=n+1}^{n+n_1} \end{array} \right\}. \qquad \mathbf{(4.16)}$$

For the truncated Poisson GLMM proposed here, Christensen et al. (2000) has shown that the LH algorithm is not geometrically ergodic because $\left\|\nabla(\gamma)\right\|$ increases very fast when $\|\gamma\| \to \infty$ in some directions. Using a truncated gradient

$$\nabla(\gamma)^{trunc} = -\gamma + \left(\Sigma^{1/2}\right)^{T} \left\{ \begin{array}{c} \left\{\left(u_i - A_i\right)\right\}_{i=1}^{n} \\ \left\{\left(v_i - \left(B_i^{*} \wedge H\right)\right)\right\}_{i=n+1}^{n+n_1} \end{array} \right\} \qquad \textbf{(4.17)}$$

where $0 < H < \infty$ is a truncation constant results in a geometrically ergodic LH algorithm for the spatial GLMMs. The binomial part of the gradient does not need to be truncated because the mean (A_i) is bounded.

In the LH update the proposal distribution is a multivariate normal distribution with mean vector $\xi(\gamma) = \gamma + \frac{h}{2}\nabla(\gamma)^{trunc}$ and covariance matrix *hI*, *h* > 0, and the acceptance probability is

$$\alpha(\gamma, \gamma') = \min\left\{ 1, \frac{f(\gamma' \mid y)\exp\left(-\frac{1}{2h}\left\|\gamma - \xi(\gamma')\right\|^{2}\right)}{f(\gamma \mid y)\exp\left(-\frac{1}{2h}\left\|\gamma' - \xi(\gamma)\right\|^{2}\right)} \right\} \qquad \textbf{(4.18)}$$

The main advantage of using LH is that it simultaneously updates the entire vector of random effects based on gradient information and can be more efficient than the fixed-scan algorithm we use in updating the regression and covariance parameters. We used this algorithm to update the random effects (**S, Z**).

In implementing the LH algorithm we encountered problems with mixing, where the proposed random effects vectors had very low acceptance rates. Roberts and Rosenthal (2001) determined 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 spatial GLMMs because the variability of individual components of the target density of interest $f\left(\mathbf{s}|\mathbf{y}\right)$ 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 $S_i \mid y_i$ will be smaller in locations with relatively higher counts. Conversely, the variance of $S_i \mid y_i$ 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 binomial case, the variance increases when the observed value approaches 0 or m_i, the number of trials at location x_i. Our application is binary, so the variance for $S_i \mid y_i$ is uniformly high for all locations.

To increase the efficiency of mixing in the presence of inhomogeneity as well as correlated components, Christensen *et al.* (2006) proposed transforming the residual vector into *a posteriori* uncorrelated components with homogeneous variance. The covariance matrix for $\mathbf{S} \mid \mathbf{y}$ is approximately $\tilde{\mathbf{\Sigma}} = \left(\mathbf{\Sigma}^{-1} + \Lambda\left(\hat{\mathbf{s}}\right)\right)^{-1}$ where $\Lambda\left(\hat{\mathbf{s}}\right)$ is a diagonal matrix with entries $-\partial^2 / \left(\partial s_i\right)^2 \log f\left(y_i \mid s_i\right), i = 1, \ldots, n,$ and $\hat{\mathbf{s}}$ is a typical value of **S**, such as the mode of

$s \mapsto f(\mathbf{s} \mid \mathbf{y})$ or the mode of $s \mapsto f(\mathbf{y} \mid \mathbf{s})$. The authors suggest updating $\tilde{\mathbf{S}}$ instead of $\mathbf{S}$ in $\mathbf{S} = \tilde{\boldsymbol{\Sigma}}^{1/2}\tilde{\mathbf{S}}$ because it has approximately uncorrelated components with homogeneous variance. In practice, it is shown that in the case of canonical links, $\Lambda(s)_{ii} = y_i$ is a good approximation in the Poisson case, and $\Lambda(s)_{ii} = y_i(1 - y_i/m_i)$ for the binomial case. In updating the random effects $(\mathbf{S}, \mathbf{Z})$ in our two-part application, we use the suggested approximation $\Lambda(z)_{ii} = y_i$ for updating **Z** in all locations with positive counts, and $\Lambda(s)_{ii} = 0$ for all locations for updating **S**.

4.5 Computing

The sampling algorithms were implemented using R (Ihaka and Gentleman, 1996), a free software environment for statistical computing and graphics (www.r-project.org). We used the computing facilities of the Pennsylvania State University High Performance Computing Group (HPC). Specifically, we used Lion-XJ, a 144-node PC computational cluster. Each node has 32 gigabytes of memory and two 3.0 GHz quad-core Intel Xeon (Woodcrest, E3450) processors. For the simulated data example consisting of 400 sample locations, it takes about 60 hours to complete 100,000 scans. In this study we generally used this number of samples to explore the properties of the posterior distribution. We expect that the computing time could be significantly reduced by using programming languages more suitable for numerical computations or by employing programming techniques that will utilize the full parallel computing capabilities of a system such as this. The computer code is archived at www.stat.psu.edu/~jlr/pub/Recta/ along with the results presented in Chapter 5.

Chapter 5

An Application to Simulated Data

Results from the application of the proposed model on simulated data are presented in this chapter. In applying the two-part model approach, we employed four covariance structures. The first is the two-stage full (TSF) covariance model described in Section **3.2**, with dependence among random effects for counts (**Z**), dependence among random effects for the binaries (**S**), and a cross-correlation among **Z** and **S**. The second is a simpler covariance with dependence among **Z** but independence among **S**, and cross-correlation among **Z** and **S**, henceforth the two stage independent binary (TSIB) model. The third takes the TSF but removes the cross-correlation among **Z** and **S**, which we call the two-stage no correlation (TSNC) model. The fourth covariance structure we studied assumes dependence among **Z**, independence among **S**, and no cross-correlation between **Z** and **S**, henceforth the two stage independent binary, no cross-correlation (TSIBNC) model.

5.1 Description of Simulation

A two-part response in 2601 (51×51 grid) equally spaced locations x_i was generated over the unit square. In each location, the two-part response (U_i, V_i) was simulated following the model described in Section 1.3: $U_i|S(x_i)$ ~ Bernoulli(A_i) and $V_i|Z(x_i)$ ~ Truncated Poisson(B_i), where

$$\begin{aligned}\operatorname{logit}(A_i) &= \alpha_0 + d(x_i)\alpha_1 + S(x_i) \\ \log(B_i) &= \beta_0 + d(x_i)\beta_1 + Z(x_i)\end{aligned} \qquad \textbf{(5.1)}$$

Conditionally on the $S(\mathbf{x})$ and $Z(\mathbf{x})$, the (U_i, V_i) are independent, and the $S(x_i)$ and $Z(x_i)$ are stationary zero-mean processes with covariances following the exponential covariance function $C(h_{ij}) = \sigma^2 \exp(-\theta h_{ij})$, where h_{ij} is the distance between locations. The cross-covariance function is constructed following Oliver (2003) as we described in Section 2.2, where $\Sigma_{SZ} = \rho_{SZ} L_S L_Z$ and L_S and L_Z are the respective Cholesky factorization matrices such that $\Sigma_S = L_S L_S^T$ and $\Sigma_Z = L_Z L_Z^T$ and $\boldsymbol{\rho}_{\mathrm{SZ}}$ is the correlation between **S** and **Z** at the same location (i.e., $\rho(S(x_i), Z(x_i)),\ i = 1, 2, \ldots$). The explanatory variable $d(x_i)$ is a function of location along the horizontal axis, $d(x_i) = 2x_i + (0.01)W_i,\ W_i \sim N(0,1)$. In all computations, $d(x)$ was centered and distances were scaled.

The true regression parameter values are $(\alpha_0, \alpha_1) = (2, 5)$ and $(\beta_0, \beta_1) = (1, 3)$ and the covariance parameters are $(\sigma^2, \theta_S, \theta_Z, \rho_{SZ}) = (1, 10, 5, 0.75)$. 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, β_1=3 means that under a constant signal z_0, the expected count (conditional on 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 θ_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 greater than one-fifth of the field. The correlation between *S* and *Z* at the same location is ρ_{SZ}=0.75, and thereafter decays exponentially with distance.

Figure **5.1** shows the simulated *S* and *Z* signals, probability of incidence $\left(E\left(U\right)\right)$, and expected counts$\left(E\left(V\right)\right)$ based on the *S* and *Z* simulated values and regression model. Both sets of random effects exhibit some spatial persistence, although the Z signals are smoother because of our choice of parameters (θ_S =10 and θ_Z= 5). There is also evident cross-correlation between them as imposed by our choice of ρ_{SZ}=0.75. Figure **5.2** shows the simulated realization of the semicontinuous variable *Y* based on the expected values of *U* and *V*. Figure **5.3** is the same plot showing only the values observed at 400 sample locations. There were 127 locations with zero counts, shown as × in the plot.

As part of an initial exploratory analysis of the data we estimated the regression parameters using a generalized linear model $g\left[E\left(\mathbf{Y}\right)\right]=\mathbf{X}\boldsymbol{\beta}$ and the appropriate link function $g\left(\cdot\right)$, assuming incorrectly that $S\left(x_i\right)=0$ and $Z\left(x_i\right)=0$ for all x. Using all sample points (including *0*s) and a log link, the estimates (and standard errors) are $\hat{\beta}_0=1.09\left(0.034\right)$ and $\hat{\beta}_1=1.60\left(0.056\right)$. Using only the positive observations, the estimates are $\hat{\beta}_0=1.48\left(0.036\right)$ and $\hat{\beta}_1=0.95\left(0.064\right)$. For the incidence model (0 vs 1) with a logit link, the estimates are $\hat{\alpha}_0=2.23\left(0.289\right)$ and $\hat{\alpha}_1=5.08\left(0.537\right)$.

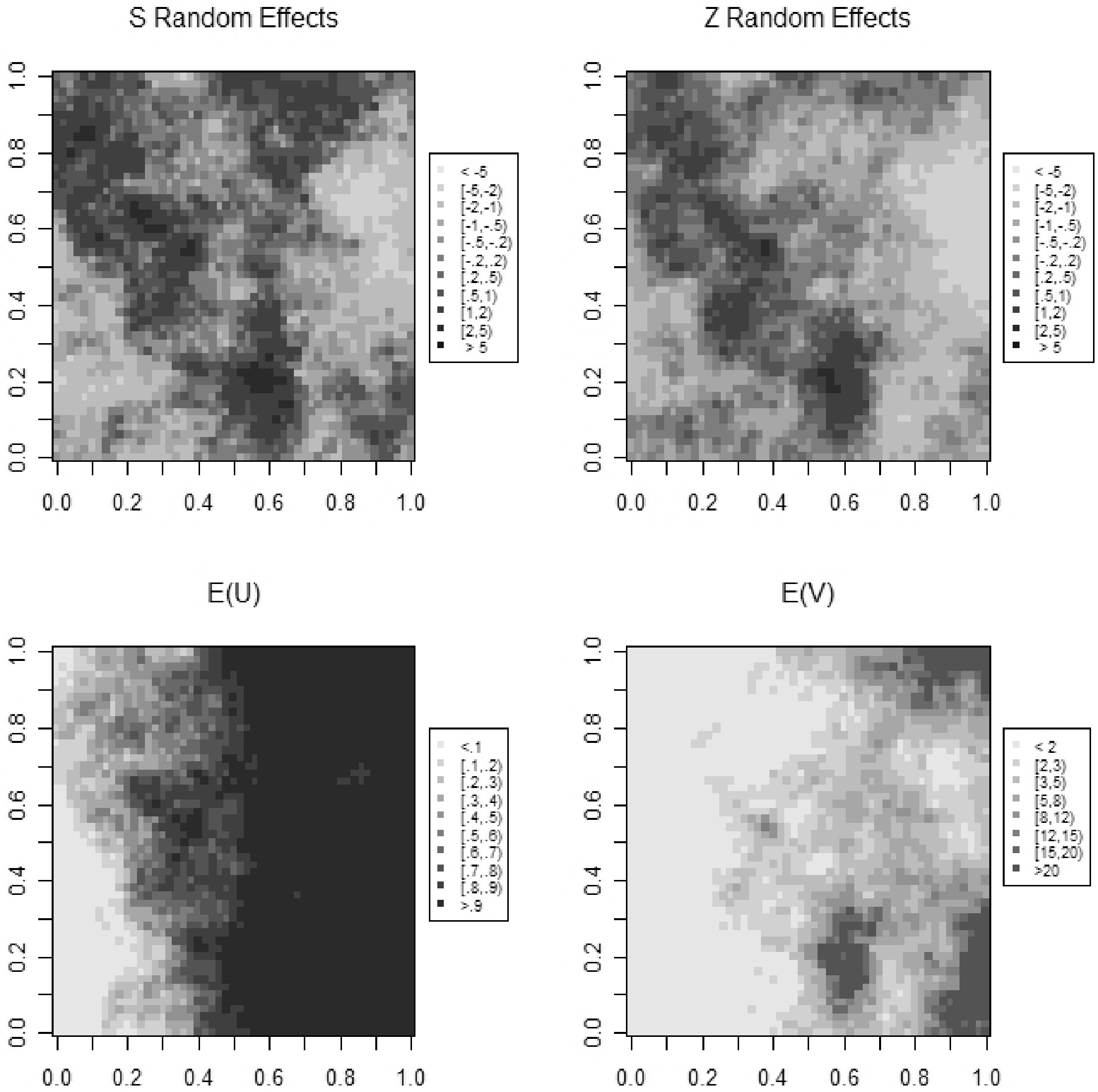


Fig. **5.1**: Simulated *S* and *Z* signals and expected values for incidence (*U*) and counts (*V*).

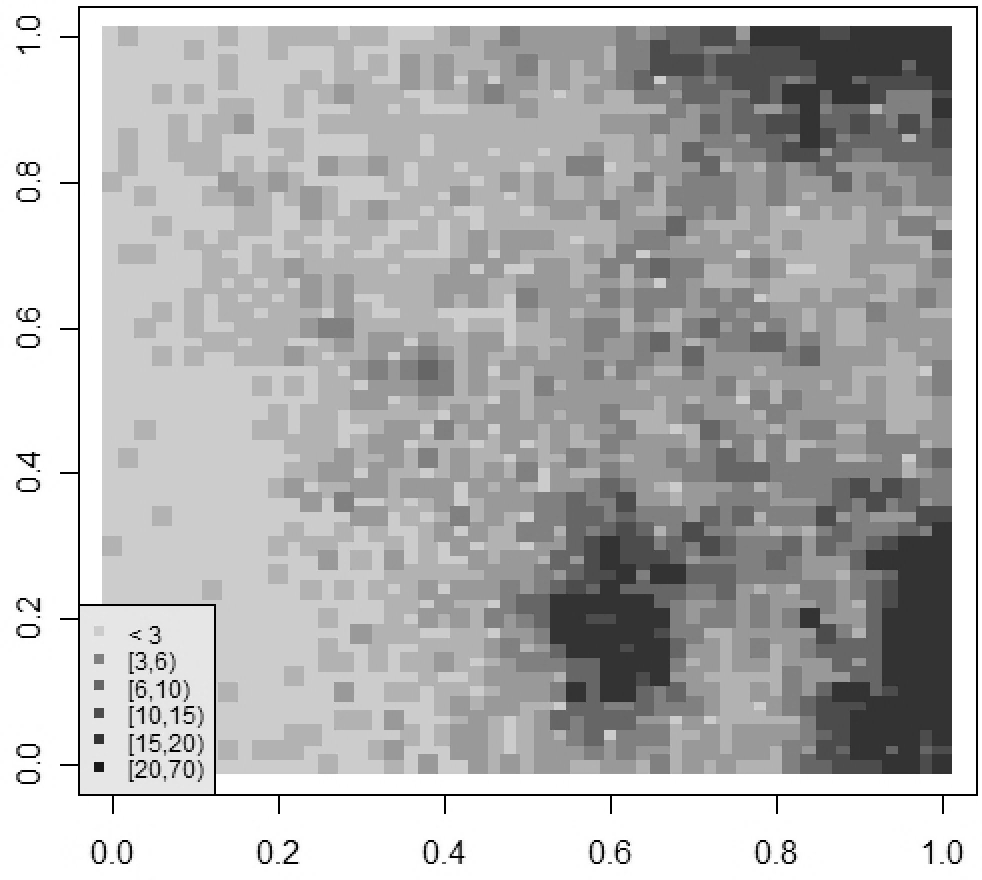


Fig. **5.2**: Semicontinuous data *Y* based on simulated values of incidence (*U*) and counts (*V*).

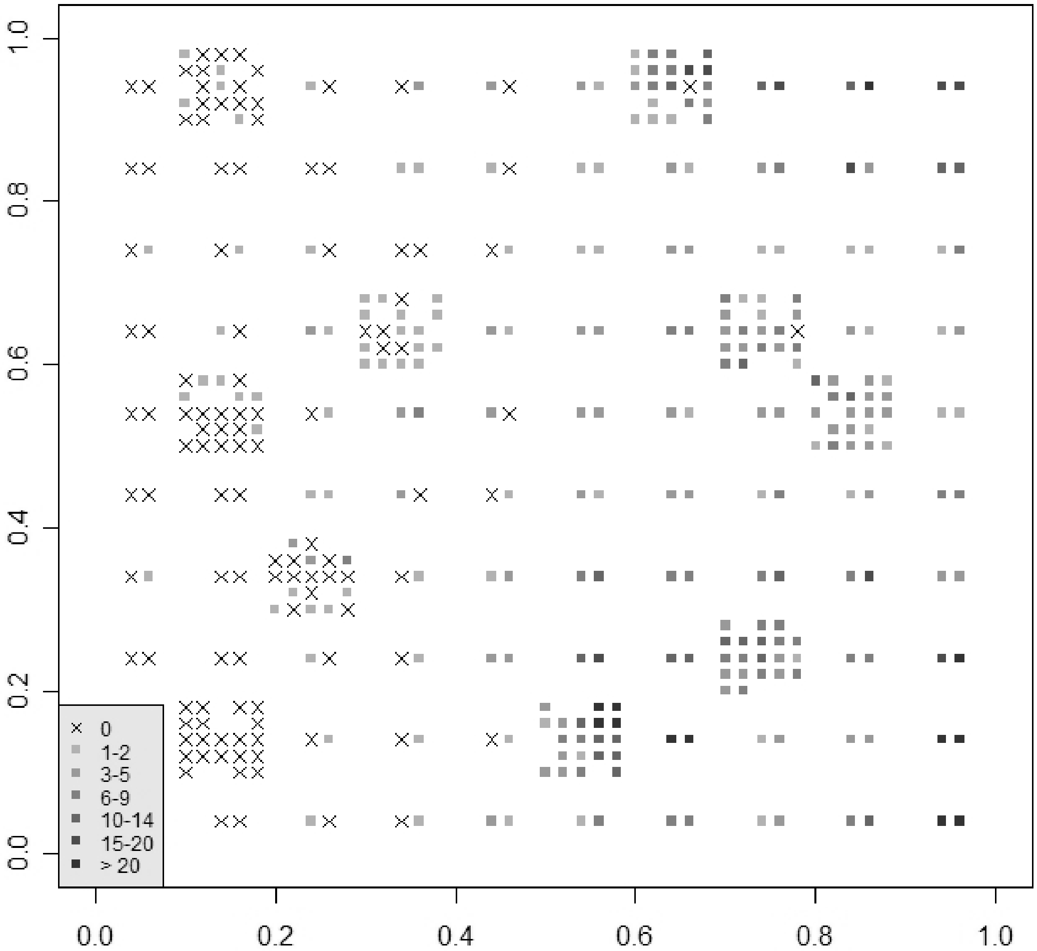


Fig. **5.3**: Observed semicontinuous data (n=400).

To estimate the spatial correlation and cross-correlation among the *S* and *Z* random effects, we grouped the pairs of observations according to distance bins and computed sample correlation and cross-correlation coefficients. Figure **5.4** plots the observed correlation coefficients as well as the theoretical value based on the specified covariance structure and parameters.

We designed a sampling plan with the objective of estimating regression coefficients as well as covariance parameters. It is important to ensure that locations are sampled throughout the field to capture mean trend, but equally important, if not more so, to ensure that there are enough sampling locations that are close together to capture the behavior of the covariance function close to the origin. Our sampling plan allocates half of the 400 locations over a regular grid and the other half to 10 clusters of 20 locations each. In this design, about 20% of the 79,800 pairs of locations have distances of 0.2 or less. Most of the information regarding spatial correlation will be obtained from these pairs because the random effects are essentially uncorrelated at distances greater than 0.2. In this subset of location pairs, about 35% will be less than 0.1 apart. From Figure **5.4** it is clear that much of the information about how fast the correlation structure decays comes from pairs of locations that are 0.1 or less apart. If sampling locations were selected at random uniformly throughout the field (i.e., no clusters), the percentage of pairs of locations with distances of 0.2 or less will be about the same (20%) but only about 25% of these pairs will be less than 0.1 apart.

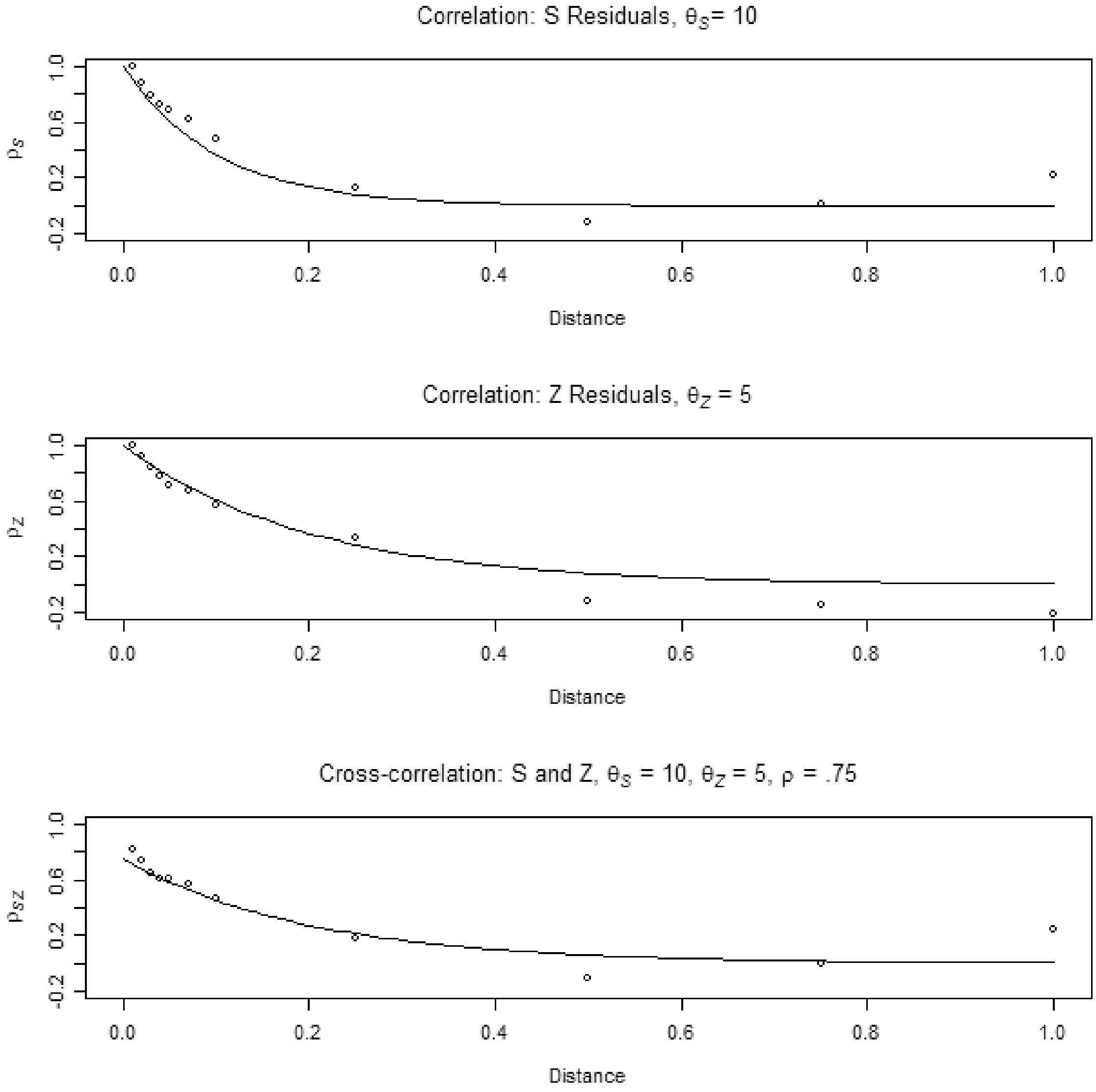


Fig. **5.4**: Observed (◦) correlation and cross-correlation coefficients computed from the sample data. The theoretical values (—) are based on the specified covariance and cross-covariance functions.

5.2 Implementation Details

In this subsection we describe details of the implementation of the MCMC algorithms described in Section **4.4**, as well as additional procedures used to summarize our findings. We use the implementation of the TSF model as an example, but the procedures described here apply to all four models.

In the MCMC implementation the initial values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are their ordinary GLM estimates, i.e., assuming independent random effects. Initial values of $\boldsymbol{\theta}$ were sampled from their uniform proper priors**,** and **S** and **Z** started as independent samples from the standard normal distribution. They can also be initially set as the residuals from the naïve GLM estimation. We sampled the posterior distributions of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$**,** and $\boldsymbol{\theta}$ using the Gaussian random walk Metropolis algorithm. We utilized log-uniform priors for θ_S and θ_Z where $\pi(\theta) \propto \theta^{-1}, \log(\theta) \in [0,5]$, a proper uniform prior for the correlation coefficient $\left(\pi(\rho) \propto 1, \rho \in [-1,1]\right)$, and flat improper priors for all regression parameters, where $\pi(\boldsymbol{\alpha},\boldsymbol{\beta}) \propto 1,\ (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \mathbb{R}^4$. For the TSIB and TSIBNC models we used a proper uniform prior for σ_S, with $\pi(\sigma) \propto 1,\ \sigma \in (0,10]$.

Posterior samples of **S** and **Z** were generated using the LH algorithm. We used truncation constant *H=30* for $\nabla(\gamma)^{trunc}$. In the LH update the proposal distribution for {**S, Z**} is a multivariate normal distribution with mean vector $\xi(\gamma) = \gamma + \frac{h}{2}\nabla(\gamma)^{trunc}$ and covariance matrix *hI*, where *h* = 0.4. We encountered problems with mixing of the distribution and implemented the reparameterization of the covariance parameters proposed by Christensen *et al*. (2006) which we summarize in Section **4.4.4**.

We describe the distribution of the samples from the MCMC procedure using posterior means and standard deviations. We also present the estimated 95% highest posterior densities (HPD) of the parameters using the approximate procedure of Chen *et al*. (2000). 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. We verified that the Monte Carlo standard errors (MCSE) for the posterior mean estimates computed by consistent batch means (Flegal *et al*., 2008; Jones *et al*., 2006) are sufficiently small. Non-overlapping batch means is one of several available methods to estimate the variance of the asymptotic distribution of a parameter of interest after generating samples from a Markov chain. This method breaks up the output of *n* values into *a* blocks of equal size *b* (so that *n*=*ab*) and computes an estimate of the variance as

$$\hat{\sigma}^2 = \frac{b}{a-1}\sum_{j=1}^{a}\left(\bar{Y}_j - \bar{\bar{Y}}\right)^2 \qquad \textbf{(5.2)}$$

where $\bar{Y}_j,\ j = 1,\ldots,a$ is the arithmetic mean in each block and $\bar{\bar{Y}}$ is the mean of all *n* values. For instance, for the TSF model (Table **5.1**) the posterior mean of β_1 is 0.99, with MCSE of 0.004. Following the procedure described in Flegal *et al.* (2008), an asymptotically valid 95% confidence interval for the expected value of β_1 is [0.98, 1.00].

The MCSEs for the parameters are presented in Table **5.1** and all similar summary tables in this chapter.

At any iteration, given the current values $\left(\boldsymbol{\theta}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\beta}^{(k)},\mathbf{S}^{(k)},\mathbf{Z}^{(k)}\right)$, we can generate samples of expected values at a set of unobserved locations **x***, which we constructed as a grid of 1600 locations. We first obtain samples $\mathbf{S}^{*}$=S(**x*)** and $\mathbf{Z}^{*}$= Z(**x*)** by noting that given the covariance parameters $\boldsymbol{\theta}^{(k)}$, $\left(\mathbf{S}^{*},\mathbf{Z}^{*},\mathbf{S}^{(k)},\mathbf{Z}^{(k)}\,|\,\boldsymbol{\theta}^{(k)}\right)$ is multivariate normal. From Eq. **4.11** and **4.12**, the distribution of $\left(\mathbf{S}^{(k)},\mathbf{Z}^{(k)}\,|\,\mathbf{U},\mathbf{V},\boldsymbol{\theta}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\beta}^{(k)}\right)$ is conditionally independent of $\left(\mathbf{S}^{*},\mathbf{Z}^{*}\right)$and samples of $\left(\mathbf{S}^{(k)},\mathbf{Z}^{(k)}\right)$ are generated during the MCMC procedure. Therefore, it is straightforward to generate samples of $\left(\mathbf{S}^{*},\mathbf{Z}^{*}\right)$from the multivariate normal distribution $\left(\mathbf{S}^{*},\mathbf{Z}^{*}\,|\,\mathbf{S}^{(k)},\mathbf{Z}^{(k)},\boldsymbol{\theta}^{(k)}\right)$. Expected values for $\mathbf{U}^{*}$ and $\mathbf{V}^{*}$ are then obtained using the specified link functions, the current values $\left(\boldsymbol{\alpha}^{(k)},\boldsymbol{\beta}^{(k)}\right)$, and the covariates $\mathbf{d}_S\left(\mathbf{x}^{*}\right)$ and and $\mathbf{d}_Z\left(\mathbf{x}^{*}\right)$. At each iteration we also generated a sample realization of $\mathbf{U}^{*}$, $\mathbf{V}^{*}$, and the semicontinuous response $\mathbf{Y}^{*}$, where $Y\left(x^{*}\right)=\begin{cases}0 & \text{if } U\left(x^{*}\right)=0\\ V\left(x^{*}\right) & \text{if } U\left(x^{*}\right)=1\end{cases}$.

We used the posterior means of $\left(\mathbf{S}^{*},\mathbf{Z}^{*},\mathbf{U}^{*},\mathbf{V}^{*},\mathbf{Y}^{*}\right)$for the prediction maps presented under each model. The program code and simulation results are archived at www.stat.psu.edu/~jlr/pub/Recta/.

5.3 Results from the TSF Covariance Model

Figure **5.5** plots MCMC samples of the covariance and regression parameters, sampling every 100th iteration after discarding the first 20000 iterations.

Table **5.1** gives the posterior mean and standard deviation of the samples, as well as estimated 95% HPD of the parameters. The posterior means of the covariance and regression parameters do not reflect the corresponding true values used to generate the data set. The degree of spatial dependence among the random effects in the simulated data was not captured in the posterior distribution, as covariance parameters were over-estimated and the cross-correlation coefficient between *S* and *Z* was close to zero. The 95% HPD for the regression parameters in the incidence part of the model ($\boldsymbol{\alpha}$) include the true values $\left(\alpha_0 = 2, \alpha_1 = 5\right)$. However, the posterior mean of the samples for β_0 (1.67) is higher than the true value of 1 while the posterior mean for slope $\left(\beta_1\right)$ for the positive counts (0.99) was underestimated; the true value is 3. We note that these posterior means compare to the maximum likelihood estimates from a naive GLM (i.e., assuming incorrectly that $Z\left(x_i\right) = 0$ for all x and V has a Poisson distribution). Using only the subset of positive observations and assuming a Poisson distribution for the observed **V** where $\log\left[E\left(\mathbf{V}\right)\right] = \mathbf{d}\left(\mathbf{X}\right)^T \boldsymbol{\beta}$, the MLE are $\hat{\beta}_0 = 1.48\left(0.036\right)$ and $\hat{\beta}_1 = 0.95\left(0.064\right)$. However we note that, as expected, the naive GLM application is overly optimistic about the precision of the estimates. The 95% confidence interval for β_1, for instance, is $\left(0.82,\ 1.08\right)$, compared to a 95% HPD $\left(0.78,\ 1.18\right)$ from the TSF model. We address the apparent bias in the estimated regression coefficients for the abundance model in the latter part of this section.

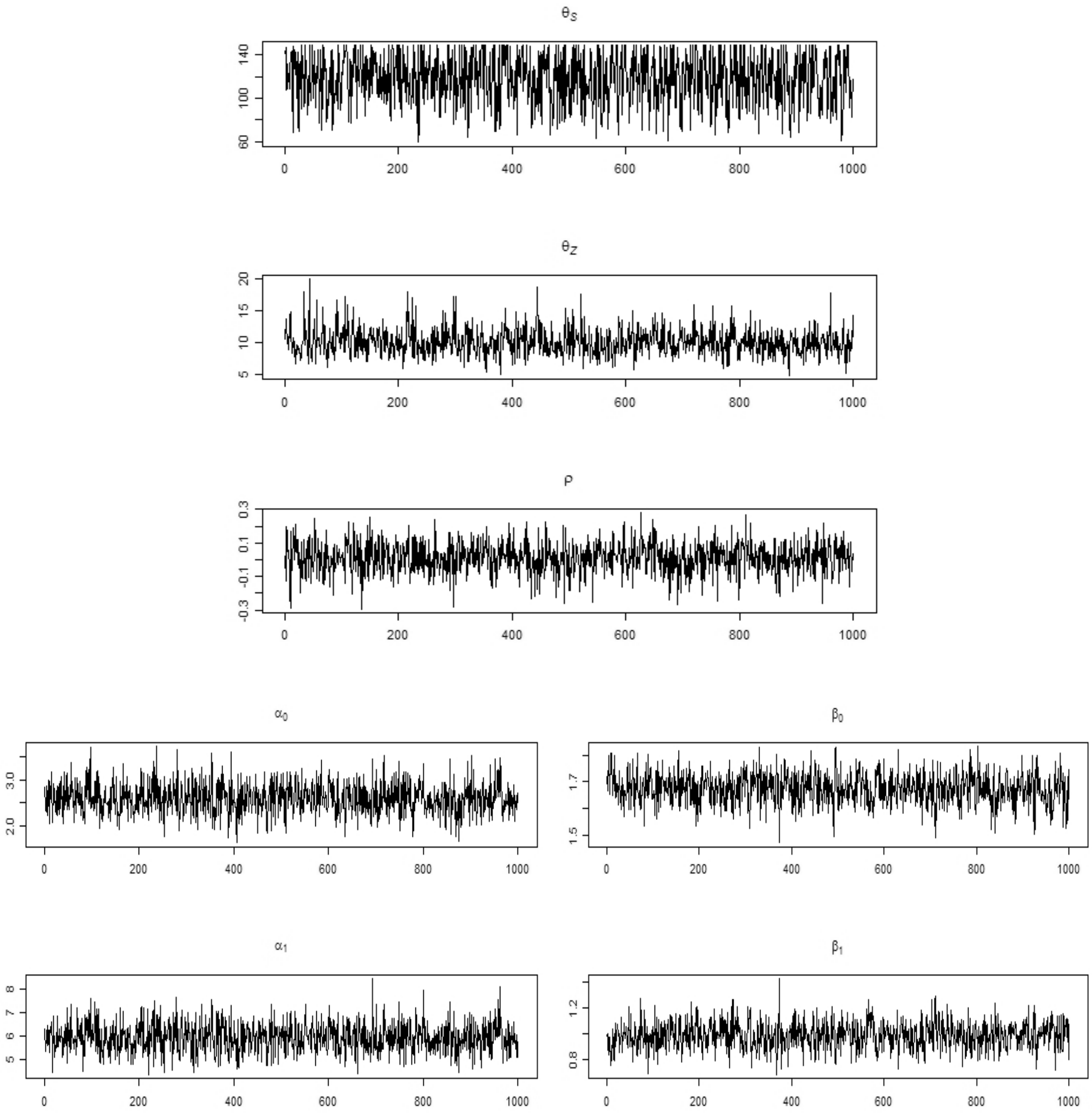


Fig. **5.5**: MCMC samples of all parameters in the TSF model.

Table **5.1**: Posterior estimates, Monte Carlo standard errors, and highest posterior densities (HPD) for parameter estimates in the simulated data set using the TSF model.

Parameter	True Value	Posterior Mean (Standard Deviation)	Monte Carlo Standard Error	95% HPD
θ_S	10.0	117.92 (23.18)	0.692	(76.79, 148.41)
θ_Z	5.0	10.03 (2.11)	0.076	(6.30, 14.27)
ρ_{SZ}	0.75	0.01 (0.09)	0.003	(-0.16, 0.21)
α_0	2.0	2.60 (0.34)	0.010	(1.95, 3.22)
α_1	5.0	5.93 (0.63)	0.018	(4.69, 7.07)
β_0	1.0	1.67 (0.06)	0.002	(1.56, 1.78)
β_1	3.0	0.99 (0.10)	0.004	(0.78,1.18)

Figure **5.6** shows the posterior mean of the **S** and **Z** samples, as well as mean predicted incidence $\left(E\left(U\right)\right)$ and positive counts $\left(E\left(V\right)\right)$. As expected from the posterior mean of 117.92 among the samples of θ_S, the *S* random effects demonstrated weak spatial dependence, but did reflect some areas with negative values found in the lower left-hand corner of the simulated data set (Figure **5.1**), showing up as lighter shades in the same areas of Figure **5.6**. These sample-based estimates of the posteriors that show lighter areas appear to be consistent with the lighter areas in the map of *Z* random effects, probably due to the cross-correlation we have included in the covariance model. The map of mean predicted incidence $\left(E\left(U\right)\right)$ shows mostly an increasing mean trend along the x-axis with few localized variations or spatial patterns corresponding to the map of *S* random effects. The samples of *Z* random effects exhibited some spatial dependence, as seen in the graph and also reflected in the mean of 10.03 among posterior samples of θ_Z (Table **5.1**). The graph of mean *Z* random effects also generally reflects the patterns seen in the simulated data set. The mean predicted positive counts $\left(E\left(V\right)\right)$ therefore shows a similar increasing mean trend along the x-axis as well as localized variation due to the spatially varying *Z* random effects that was detected by the posterior distribution of this model. The posterior distribution did not indicate a correlation between *S* and *Z*, with a posterior mean of 0.02 for the ρ, and a 95% HPD of (-0.16, 0.21), which contains 0.

At each iteration and prediction location, we generated a realization *U* and *V*, and hence the semicontinuous variable *Y*, where Y=0 when *U*=0 and $Y=V$ otherwise. The posterior means of the *Y* and the 95% HPD for the predicted *Y* are mapped in Figure **5.7**. This map is consistent with the simulated values in Figure **5.2** , generally increasing along the x-axis with some areas of higher abundance that reflect clusters of higher Z random effects at the top right and bottom right corners.

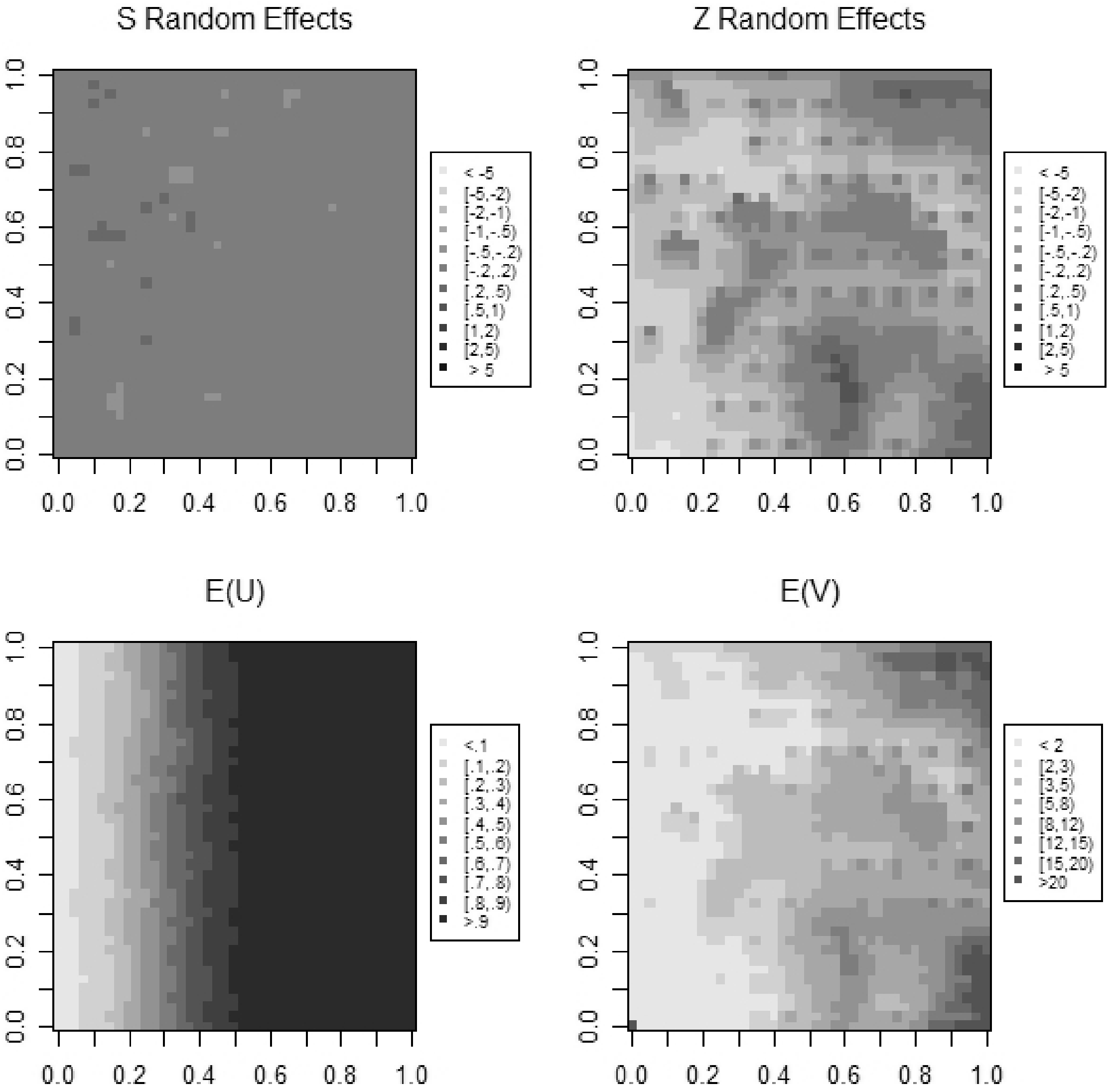


Fig. **5.6**: Predicted mean surface for simulated data using the TSF model.

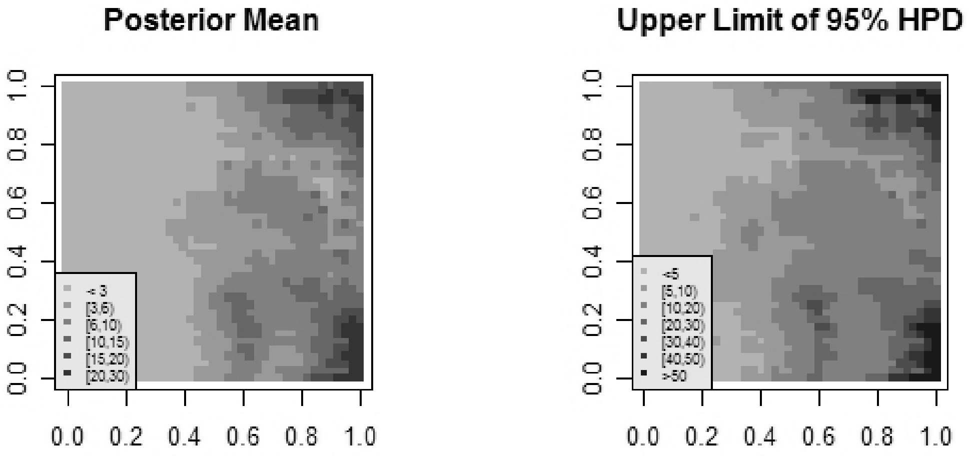


Fig. **5.7**: Posterior mean and 95% HPD for the semicontinuous response *Y* using the TSF model. The lower limit of the 95% HPD is <5 for 99% of all locations and is not mapped here.

The difficulty of capturing spatial dependence in a spatial logistic mixed model setting has been observed in previous work. Liang *et al*. (2008) found that the correlation parameters were only weakly identified by the data, leading to poor MCMC convergence. They set the spatial correlation parameters to selected values and used them as tuning constants in order to study other parameters which were of greater interest, such as the regression coefficients. These results are also consistent with the conclusions based on theory and simulation in Zhang (2004) who found, in the context of maximum likelihood inference for model-based geostatistics, that in model-based geostatistics, not all parameters in the Matérn class (which includes the exponential covariance function used here) can be estimated consistently even if data are observed in an increasing density over a fixed domain.

In the binary case where the only observation in a location x_i is incidence, there is some indication on whether the spatial residual S_i is positive or negative, but very little information regarding its magnitude. If $U(x_i)$=1, large positive proposal values for S_i for the Gaussian quantity $\text{logit}(A_i) = \mathbf{d}_S(x_i)^T \boldsymbol{\alpha} + S(x_i)$ will be more likely to be accepted. Conversely, when $U(x_j)$=0, large negative proposal values for S_i will be accepted more often.

Earlier in this section we observed that the posterior mean of the samples for β_0 (1.67) is higher than the true value of 1 while the posterior mean for slope (β_1) for the positive counts (0.99) was underestimated; the true value is 3. We also note that these posterior means are similar to estimates from a naive GLM application on the positive observations, where $\hat{\beta}_0 = 1.48$ and $\hat{\beta}_1 = 0.95$.

The point estimate from the GLM is close to the marginal posterior mean for β_1 due to the particular realization of Z in this data set. Figure **5.8** (a) plots the actual simulated Z values in the data set against the $d(x_i)$, the value of the covariate. The locations where Y>0 (equivalently, U=1) are plotted as solid circles (•) and × shows locations where we have a realization Z_i in the data set but no observed V_i because the U_i=0; therefore we do not have a direct estimate of Z_i in these locations. This particular realization of our assumed Gaussian process shows a negative association with the lone covariate $d(x_i)$, particularly among Y>0. A least squares estimate of the linear relationship between Z_i and $d(x_i)$ in this subset of locations gives an intercept of -0.02 and slope of -1.58.

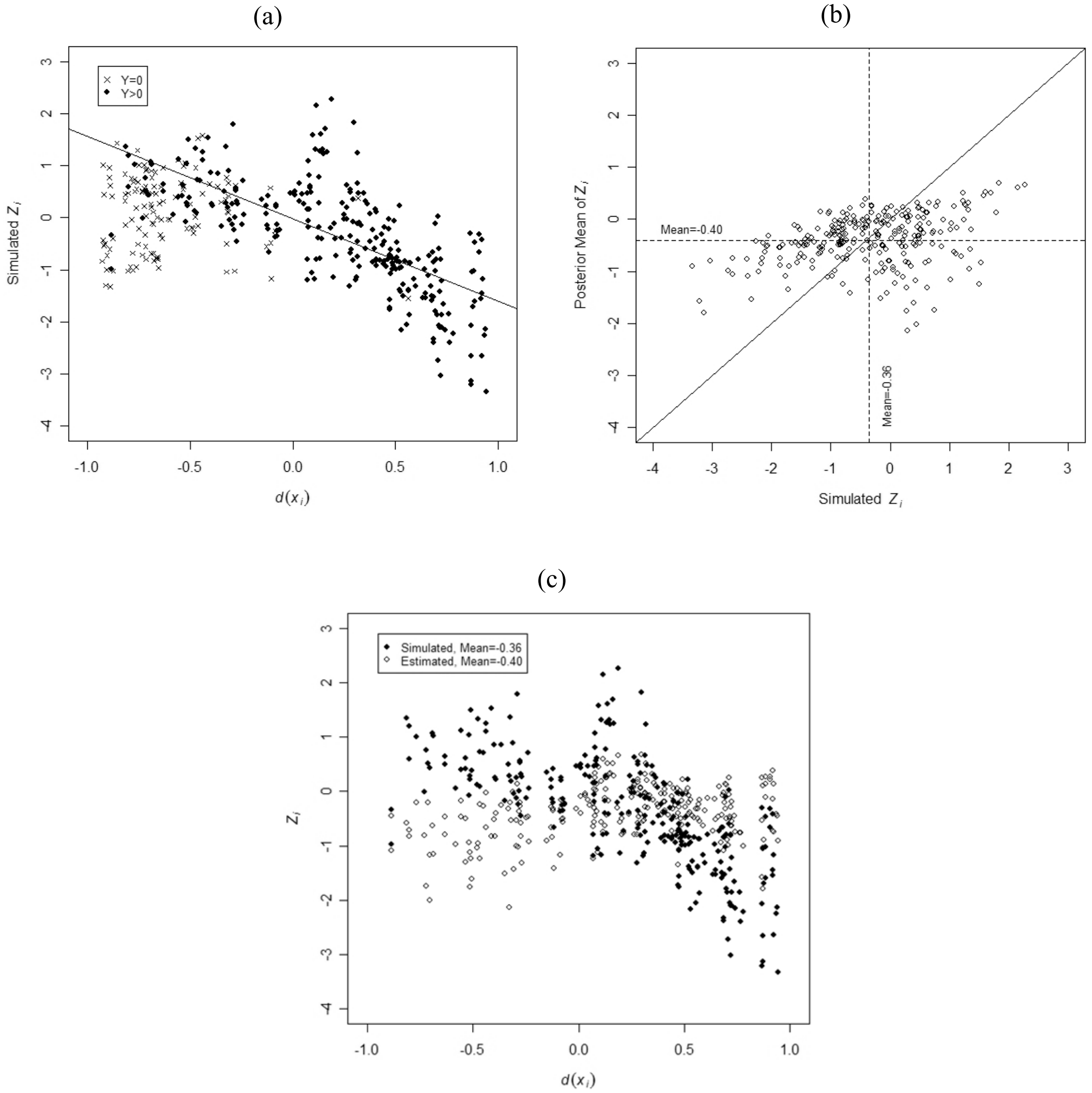


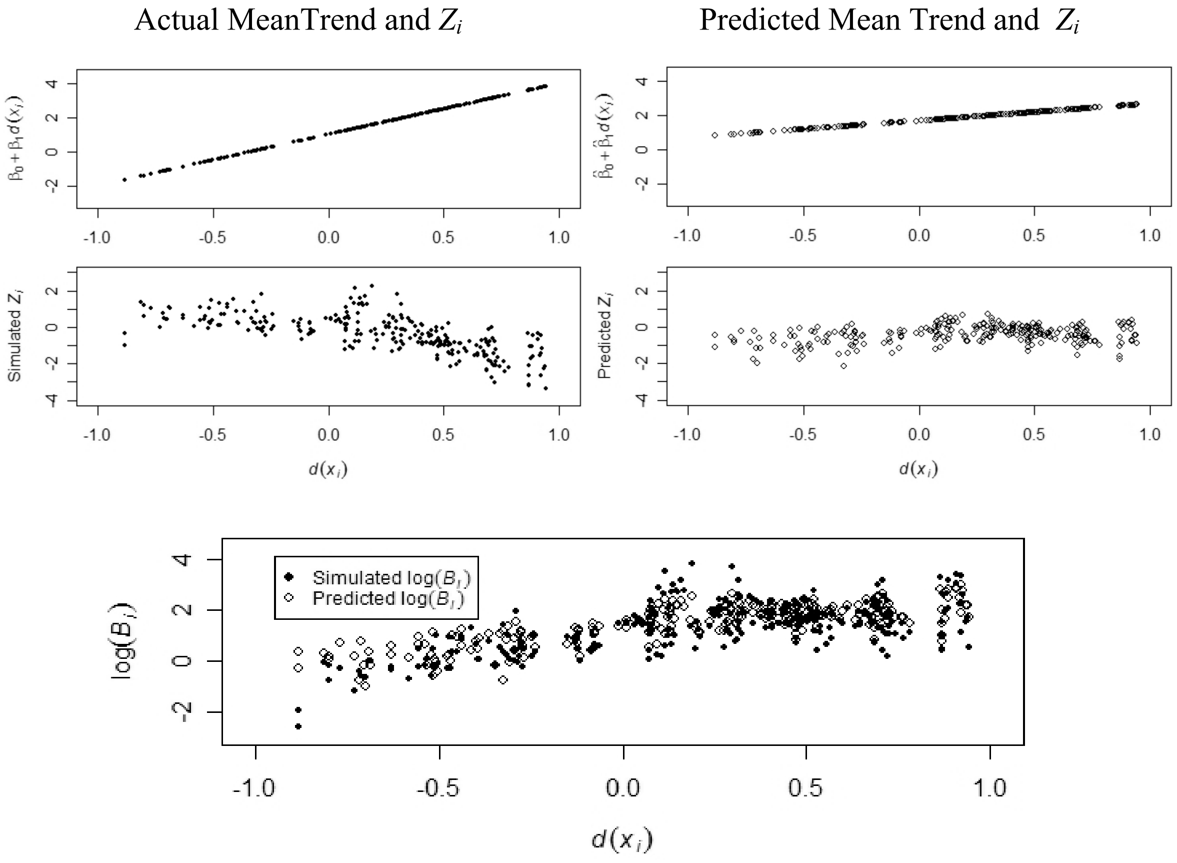
Fig. **5.8**: Plots of the Z_i random effects. (a) Simulated Z_i vs. $d(x_i)$ with ordinary least squares regression line $E\left(Z_i \middle| Y_i > 0\right) = -0.02 - 1.58d\left(x_i\right)$; (b) Simulated Z_i vs. corresponding mean posterior from TSF model; (c) Simulated and mean posterior Z_i vs. $d(x_i)$

Figure **5.8** (b) plots the same simulated (actual) *Z* values in locations where abundance is observed (*Y*>0) against their corresponding posterior means under the TSF model. Although both surfaces are centered on the same value (-0.40 vs. -0.36), the posterior mean values for Z_i are smoother than the actual values, with no extreme high or low values. Without these high and low values there is no apparent association between $d(x_i)$ and the posterior Z_i as shown in Figure **5.8** (c).

The posterior random effects are smoother than the actual Z_i due to two related reasons. First, the decreasing trend among the actual Z_i along $d(x_i)$ has been attributed to the lone covariate in the model, which is $d(x_i)$. Generally, the positive slope $\beta_1 = 3$ was countered by the spurious negative trend in Z, resulting in a lower slope $\hat{\beta}_1 = 0.99$. Secondly, this "attribution" is also a consequence of the covariance model we impose on the random effect *Z*. The spatial dependence imposed on the random effects favors a smoother random effects surface, where Z_i closer together in space are more alike, and large differences among adjacent Z_i will be highly unlikely.

Figure **5.9** graphically illustrates this confounding between the mean trend and random effect Z. The left column shows the mean trend $\left(1+3d\left(x_i\right)\right)$ and Z_i in the simulated data set, where the mean trend increases with $d\left(x_i\right)$ but the opposite is true for Z_i , and the resulting $\log\left(B_i\right)=1+3d\left(x_i\right)+Z_i$ are plotted as solid circles in the bottom graph, where $V_i \sim$ Truncated Poisson(B_i). The right column shows the mean trend $\left(1.67+0.99d\left(x_i\right)\right)$ and Z_i using the posterior means from the TSF model, showing a more modest mean trend and smoothly varying Z_i, which provided good estimates of $\log\left(B_i\right)$

plotted as hollow circles in the bottom graph. We should note, however, that these graphs only illustrate the intuition behind this phenomenon in the linear part of the model, as we do not have closed form expressions of how individual values of random effects and their relationship to fixed effects can change the estimated coefficients in the spatial GLMM setting.


Actual MeanTrend and Z_i

Predicted Mean Trend and Z_i

Fig. **5.9**: Mean trend and random effects in the simulated data set (left column) and from the posterior distributions in the TSF model (right column), and their sum log(B_i).

In the case of a CAR model with normally distributed observations, Reich *et al*. (2006) show that the posterior mean and variance of fixed effects can differ substantially between a non-spatial regression approach and after accounting for spatial correlation. In particular, they observe that as spatial correlation increases, "random effects are smoothed to zero and β's posterior is similar to its posterior under the ordinary linear model." In our case, the random effects for the abundance part were smoothed to zero because the trend was attributed to the covariate, resulting in spatial regression coefficients that are close to the naive GLM estimates.

Based on these findings it is important to emphasize that the regression parameters must be interpreted conditionally on the random effects, rather than marginally. The need to distinguish between conditional and marginal regression parameters, which does not arise in linear Gaussian models, is well known in the context of GLMMs for longitudinal data (see for example Diggle *et al*., 1994, Chapter 7), and, in this case, equally so for spatial GLMMs.

To further demonstrate the confounding between random effects and regression parameters in the case of generalized linear mixed models, we confirm that we can recover the correct regression parameters if we apply the above MH algorithms to sample the covariance and regression parameters while fixing all *S* and *Z* random effects at their (known) simulated values. Table **5.2** summarizes the posterior samples taken under this constraint. All parameters were appropriately captured when the uncertainty from the *S, Z* random effects is removed. Clearly this scenario has no practical utility; it is only intended to verify that the posterior distributions for the parameters in the fixed part of the model are being sampled appropriately and to demonstrate the conditional interpretation of the regression parameters in light of random effects.

Table **5.2**: Posterior estimates, Monte Carlo standard errors, and HPD for parameter estimates in the simulated data set when the (**S, Z**) are fixed at their simulated values, using the TSF model.

Parameter	True Value	Posterior Mean (Standard Deviation)	Monte Carlo Standard Error	95% HPD
θ_S	10.0	9.70 (0.75)	0.024	(8.36, 11.23)
θ_Z	5.0	5.00 (0.43)	0.012	(4.23, 5.83)
ρ_{SZ}	0.75	0.74 (0.03)	9×10^{-4}	(0.70, 0.80)
α_0	2.0	2.08 (0.29)	0.010	(1.49, 2.59)
α_1	5.0	5.57 (0.54)	0.009	(4.64, 6.74)
β_0	1.0	0.97 (0.04)	0.001	(0.88,1.05)
β_1	3.0	3.06 (0.08)	0.003	(2.90,3.21)

Figure **5.10** presents histograms of the posterior samples for the covariance and regression parameters.

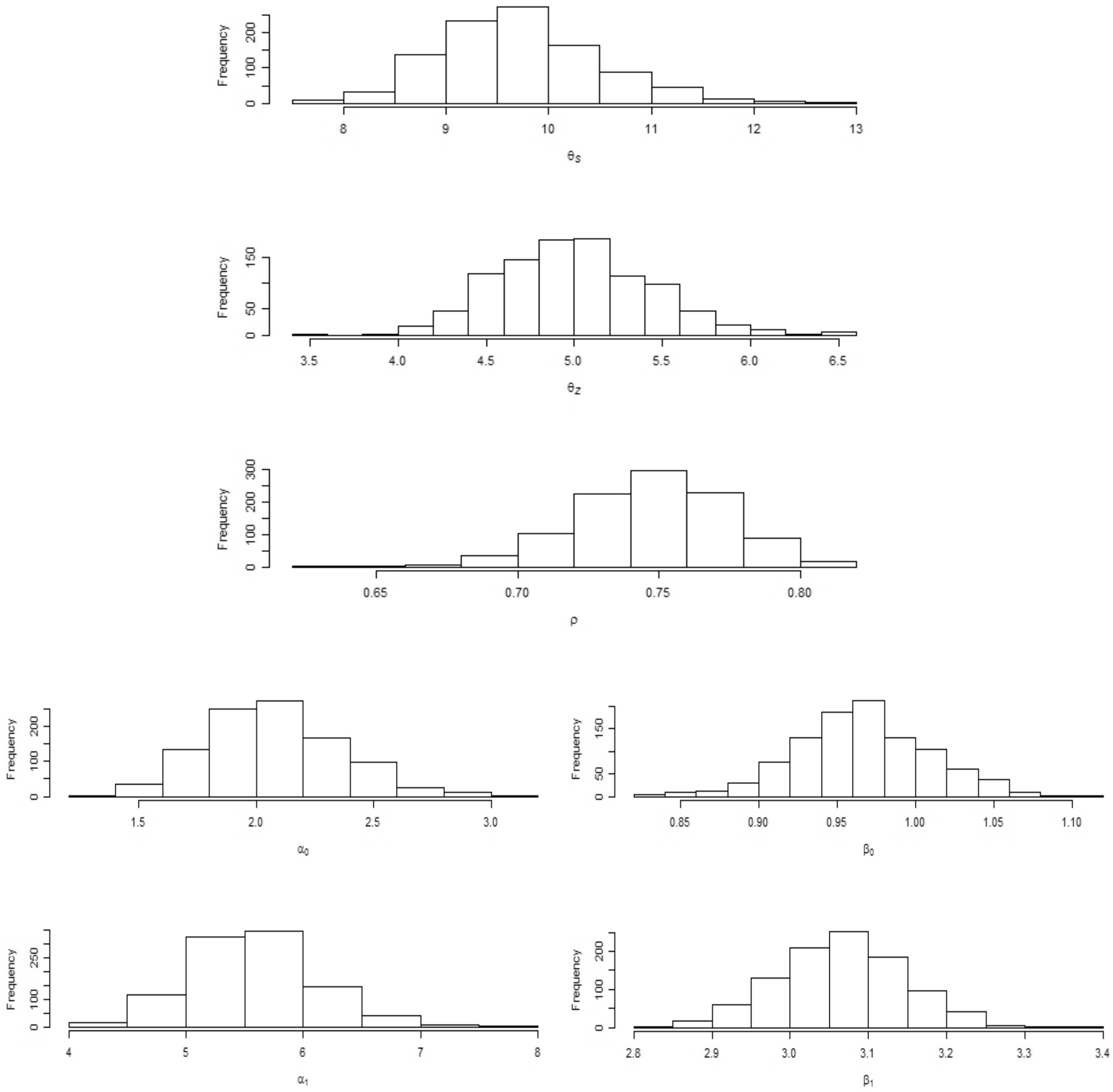


Fig. **5.10**: Distribution of posterior samples of all parameters for the simulated data set when using the TSF model and fixing (**S, Z**) at their simulated values.

We also performed the same sampling algorithms with the covariance parameters fixed at their true values, i.e. $\theta_S = 10,\ \theta_Z = 5,$ and $\rho_{SZ} = 0.75$. The results are summarized in Table **5.3**. The resulting posterior means for the regression coefficients are not unlike those from the full TSF model in Table **5.1** for the same reasons as in the

TSF model. By imposing a priori strong spatial dependence among the random effects, the collinearity between the actual values of the *Z* and the lone regressor *d*(*x*) will be reflected in the regression coefficient for *d*(*x*) and smoothed random effects.

Table **5.3**: Posterior estimates, Monte Carlo standard errors, and HPD for parameter estimates in the simulated data set using the TSF model when the covariance parameters are fixed at their known values

Parameter	True Value	Posterior Mean (Standard Deviation)	Monte Carlo Standard Error	95% HPD
α_0	2.0	2.49 (0.34)	0.011	(1.82, 3.15)
α_1	5.0	5.37 (0.63)	0.016	(4.12, 6.54)
β_0	1.0	1.55 (0.05)	0.001	(1.46,1.66)
β_1	3.0	1.05 (0.09)	0.002	(0.86,1.22)

5.4 Results from the TSIB Model

In this subsection we explore whether it is possible to indirectly establish spatial dependence among **S** through its correlation with **Z**, and what effect using a simpler covariance structure will have on the parameter estimates as well as prediction. In TSIB, we employ a simpler covariance structure, simply setting **S** as a vector of independent normal random effects. We keep all elements described in the preliminary model in Section **3.2** except for the exponential covariance structure for **S.** Instead, we have

$$\begin{aligned} \operatorname{cov}(S(x_i), S(x_j)) &= \sigma_{\mathrm{S}}^2 I_n \\ \operatorname{cov}(Z(x_i), Z(x_j)) &= \sigma_{\mathrm{Z}}^2 \exp\left[-\theta_{\mathrm{Z}} \left\| x_i - x_j \right\|\right] \end{aligned} \qquad \mathbf{(5.3)}$$

for some $\sigma_S^2 > 0,\ \sigma_Z^2 > 0$ and $\theta_Z > 0.$ We constructed the cross-covariance function using the procedure of Oliver (2003) which we described in Section **2.3**. In this formulation, $\Sigma_{SZ} = \rho_{SZ} L_S L_Z^T,$ where $\boldsymbol{\rho}_{SZ}$ is the correlation between **S** and **Z** at the same location and L_S and L_Z are the respective Cholesky factors from $\Sigma_S = L_S L_S^T$, and $\Sigma_Z = L_Z L_Z^T$.

We implemented MCMC using the procedures described in Section **5.2**.

Figure **5.11** has plots of the posterior samples of the covariance and regression parameters. The covariance parameters, particularly θ_Z andρ_{SZ}, appear to be comparable to the corresponding MCMC samples obtained under the TSF model (Figure **5.5**) in the previous section. The sample values for the regression coefficients appear to be centered around values higher than their corresponding true values, except for β_1 samples, which has a posterior mean that is lower than its true value (Table **5.4**).

Table **5.4**: Posterior estimates, Monte Carlo standard errors, and HPD for parameter estimates in the simulated data using the TSIB model.

Parameter	True Value	Posterior Mean (Standard Deviation)	Monte Carlo Standard Error	95% HPD
σ_S	1.0	5.08 (1.08)	0.176	(3.12, 6.88)
θ_Z	5.0	17.98 (5.23)	0.496	(7.49, 27.38)
ρ_{SZ}	0.75	-0.02 (0.10)	0.007	(-0.20, 0.16)
α_0	2.0	6.43 (1.22)	0.169	(4.00, 8.65)
α_1	5.0	15.13 (2.86)	0.430	(9.71, 20.27)
β_0	1.0	1.62 (0.06)	0.004	(1.51, 1.74)
β_1	3.0	1.03 (0.10)	0.007	(0.84, 1.24)

Figure **5.12** shows the predicted surfaces for both sets of spatial random effects, as well as mean incidence $\left(E\left(U\right)\right)$ and mean positive$\left(E\left(V\right)\right)$ counts. By construction, the *S* random effects are uncorrelated, and this is reflected in the apparent lack of spatial dependence or clustering in the predicted surface. Table **5.4** also shows no correlation between *S* and *Z*, so that the posterior distribution was not able to draw on any correlation between the two Gaussian processes to approximate the spatial structure of the *S* based on the observed spatial dependence among the *Z*. The map of mean incidence $\left(E\left(U\right)\right)$ is then mostly an increasing function of distance from the origin of the X-axis.

The *Z* random effects exhibited weak spatial dependence, showing some of the patterns that exist in the simulated data. These patterns carry over to the predicted positive $\left(E\left(V\right)\right)$ count surface map, which reflects the same patterns as the simulated data set in Figure **5.1**.

Figure **5.13** maps the posterior mean and HPD of the expected value of the semicontinuous variable *Y*, as we did in Figure **5.7** for the TSF model. This map is consistent with the simulated values in Figure **5.2** , generally increasing along the x-axis with some areas of higher abundance that reflect clusters of higher Z random effects at the top right and bottom right corners.

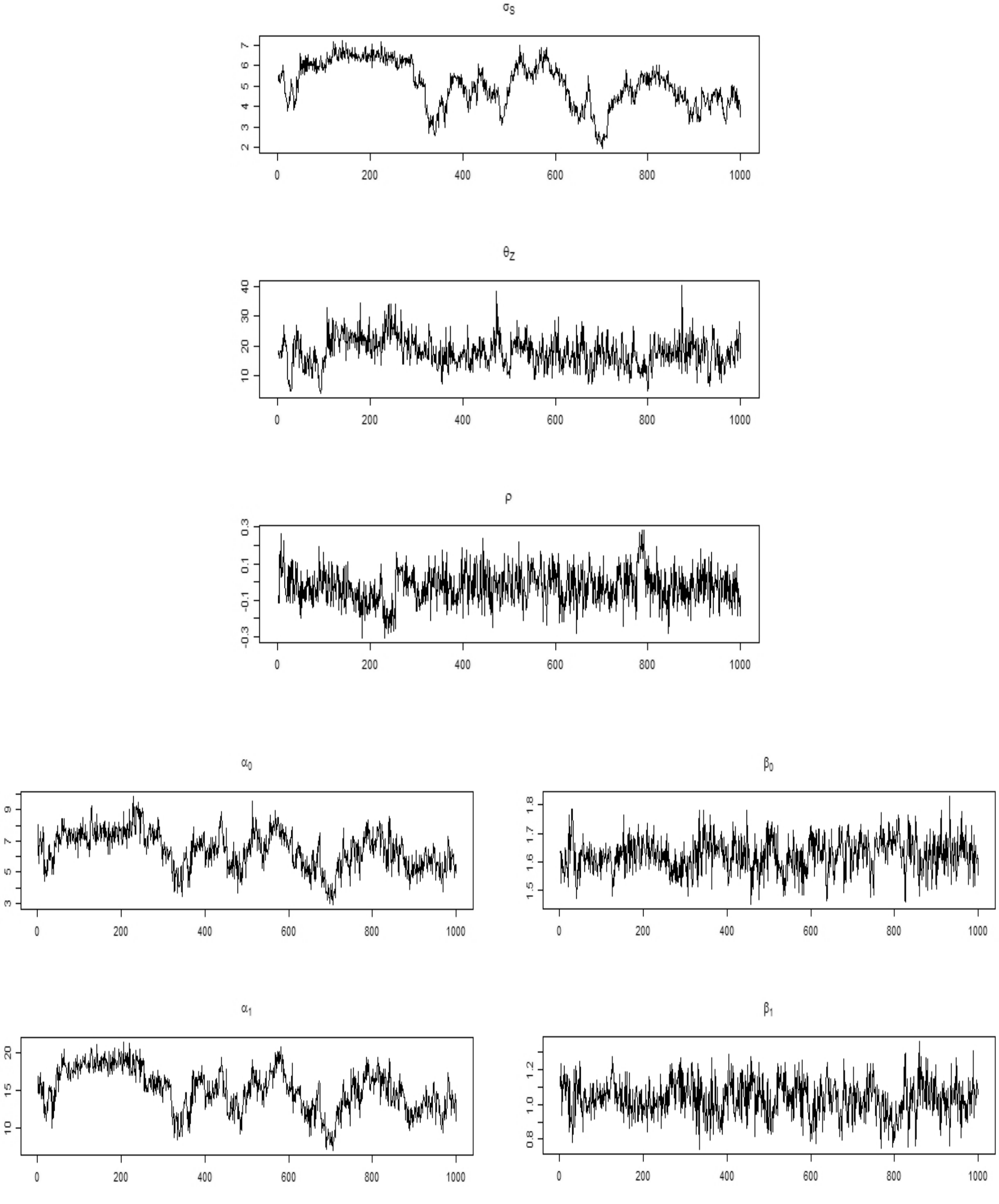


Fig. **5.11**: MCMC samples of all parameters in the TSIB model.

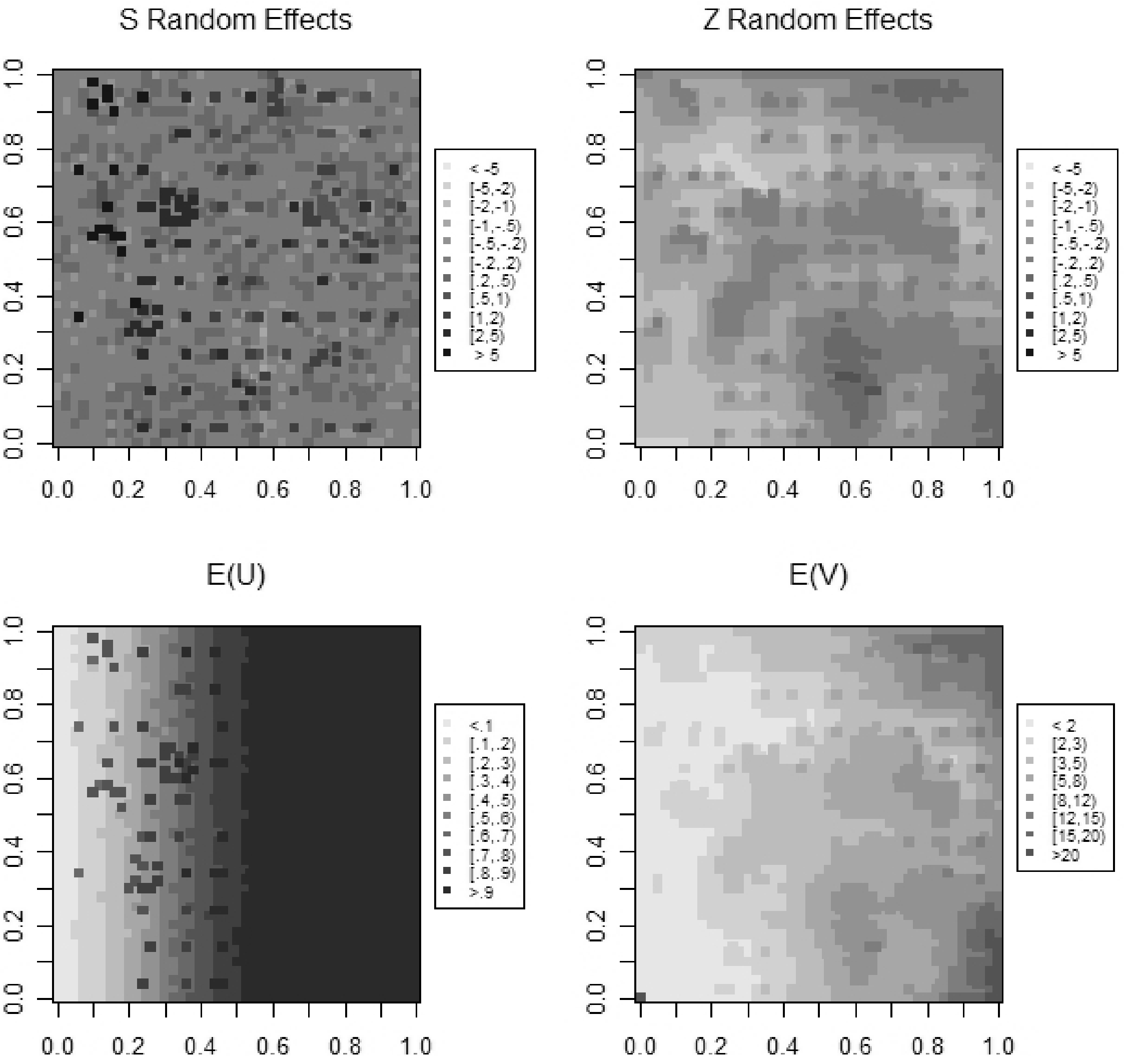


Fig. **5.12**: Predicted mean surface for simulated data using the TSIB model.

We simplified the covariance structure such that we no longer attempt to capture the spatial dependence among the *S* random effects. This also simplified the cross-covariance between *S* and *Z*, although we purposely kept the correlation between the two sets of random effects to find out if the spatial structure captured among the *Z* can induce a similar spatial structure in the *S* process. We had hoped that the *Z*s would essentially rein in the *S* random effects by virtue of the correlation between these two processes. Unfortunately, it appears that the apparent lack of spatial structure among the *S* random effects cannot be overcome by the observed spatial dependence in a related process.

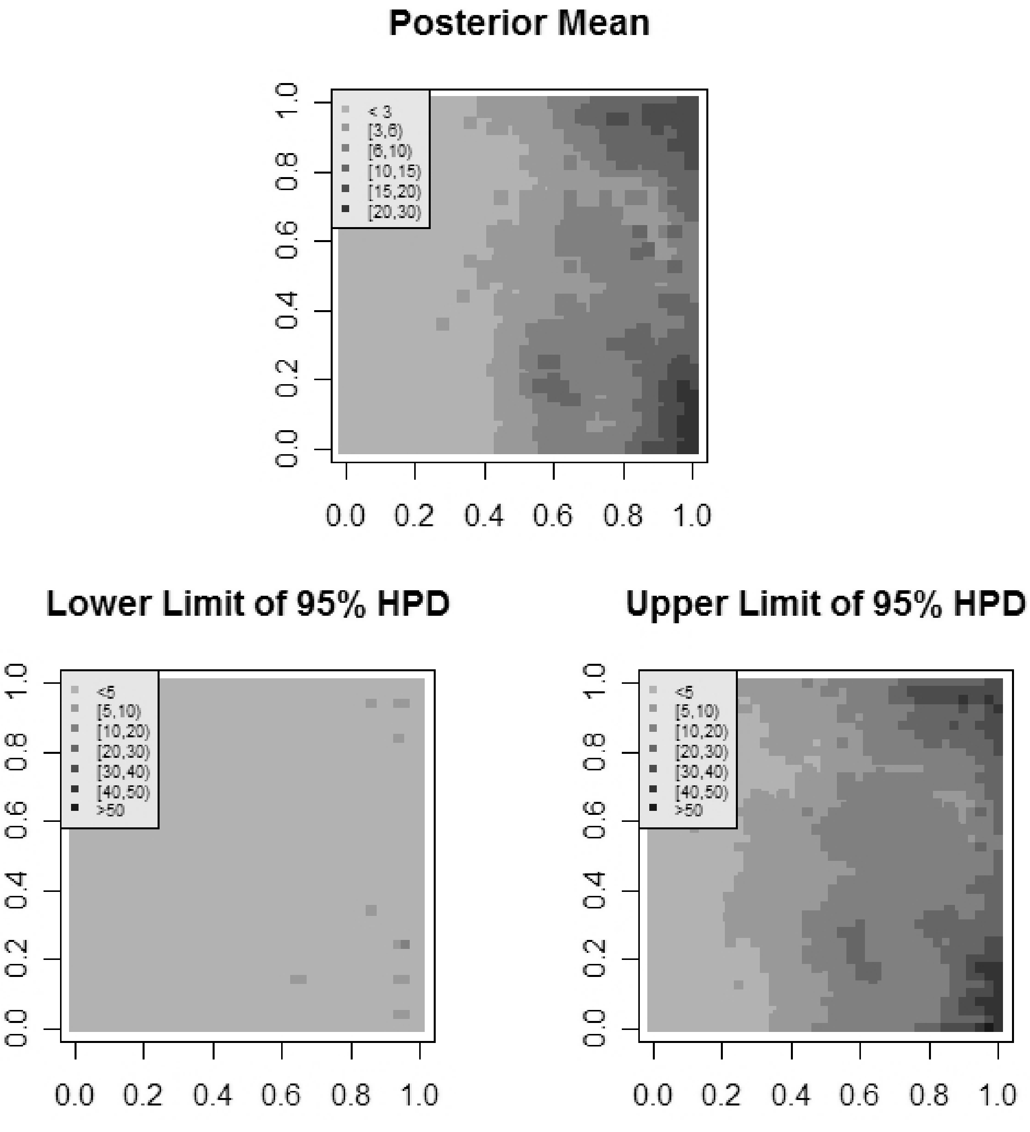


Fig. **5.13**: Posterior mean and 95% HPD for the semicontinuous response *Y* using the TSIB model.

5.5 Results Using the TSNC Model

The TSNC model assumes spatially dependent covariance functions for each set of random effects but removes the cross-correlation between these sets. This means we kept all elements of the TSF model but set $\rho_{SZ} = 0$. We implemented MCMC using the algorithms described in Section 4.4 , using the same priors as the TSF model as detailed in Section **5.2**.

Figure **5.14** plots the posterior samples for all parameters and Table **5.5** summarizes the sample values from the posterior distribution of the TSNC model. The posterior samples of the covariance parameters appear to be comparable to those from the TSF model. This is not surprising because the HPD for ρ_{SZ} in the TSF model contains 0, while we fixed it at 0 in the TSNC model. The posterior samples for regression coefficients in the abundance ($\boldsymbol{\beta}$) part of the model are also quite similar to those in the TSF, for the same reason. The 95% HPD for the regression coefficients in the incidence part contain the true values, α_0 =2 and $\alpha_1 = 5$.

Table **5.5**: Posterior estimates, Monte Carlo standard errors, and HPD for parameter estimates in the simulated data using the TSNC model.

Parameter	True Value	Posterior Mean (Standard Deviation)	Monte Carlo Standard Error	95% HPD
θ_S	10.0	140.85 (17.30)	0.513	(94.83, 148.41)
θ_Z	5.0	10.27 (2.31)	0.060	(6.34, 15.01)
α_0	2.0	2.60 (0.33)	0.011	(1.94, 3.22)
α_1	5.0	5.91 (0.61)	0.018	(4.71, 7.05)
β_0	1.0	1.67 (0.06)	0.002	(1.56, 1.79)
β_1	3.0	0.98 (0.11)	0.003	(0.78, 1.19)

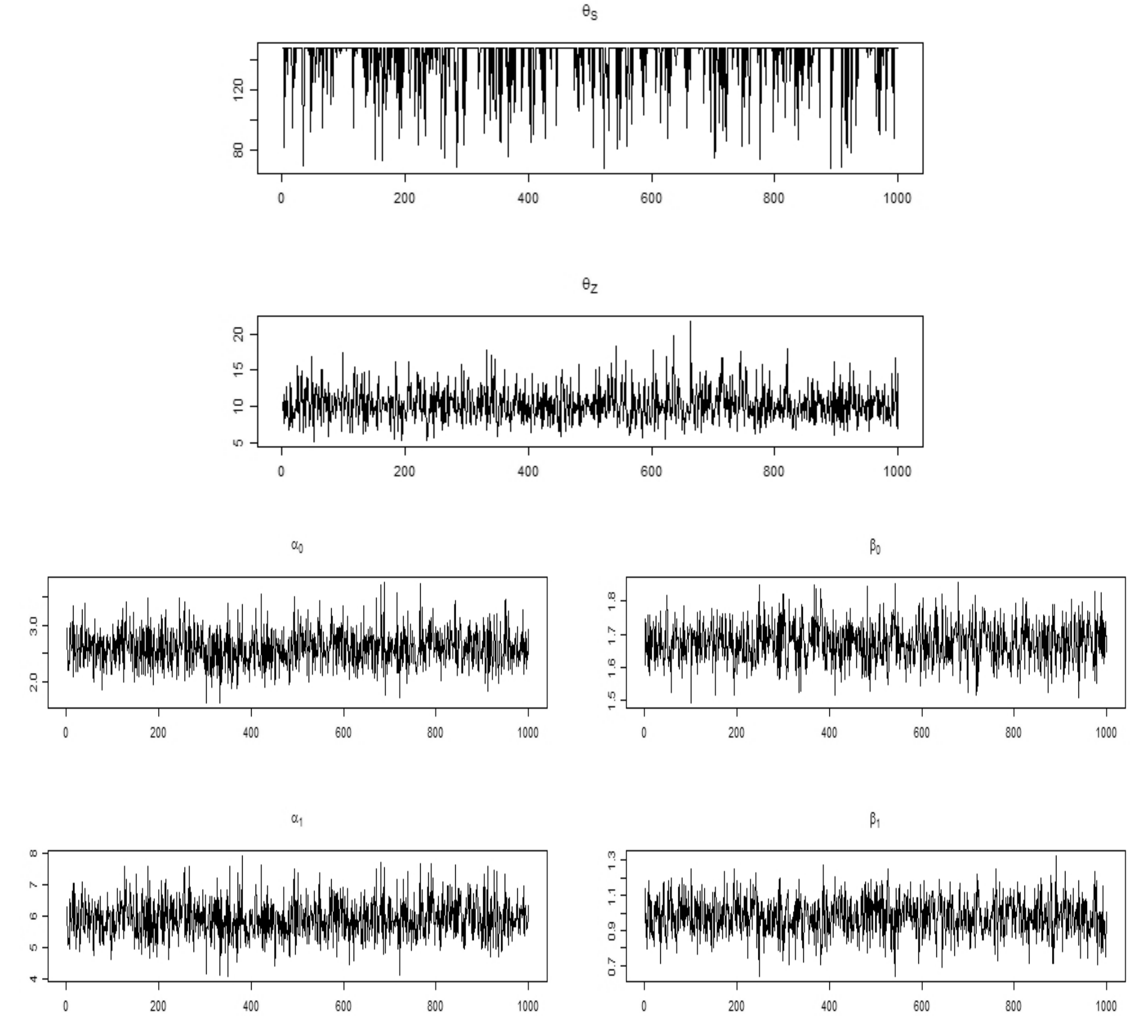


Fig. **5.14**: MCMC samples of covariance and regression parameters under the TSNC model.

Figure **5.15** maps the predicted surfaces for both sets of spatial random effects, as well as mean incidence $\left(E(U)\right)$ and mean positive$\left(E(V)\right)$ counts, and Figure **5.16** maps the posterior mean and upper limit of the HPD of the semicontinuous variable *Y*. These maps are very similar to the corresponding maps in the TSF model, as we have seen that the posterior distribution under the TSF model indicated absence of spatial dependence among the *S* random effects as well as absence of cross-correlation between the **S** and **Z**.

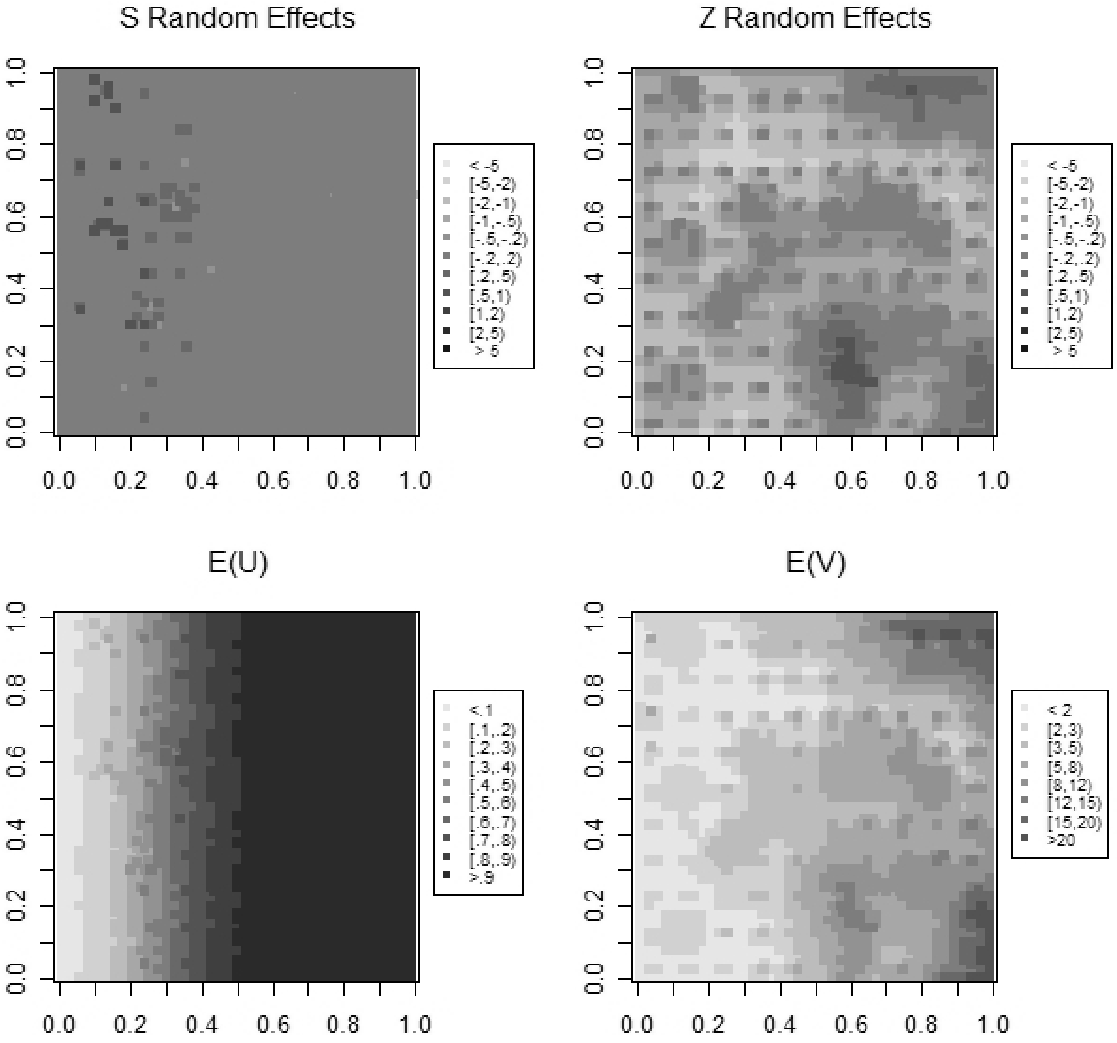


Fig. **5.15**: Predicted mean surface for simulated data using the TSNC model.

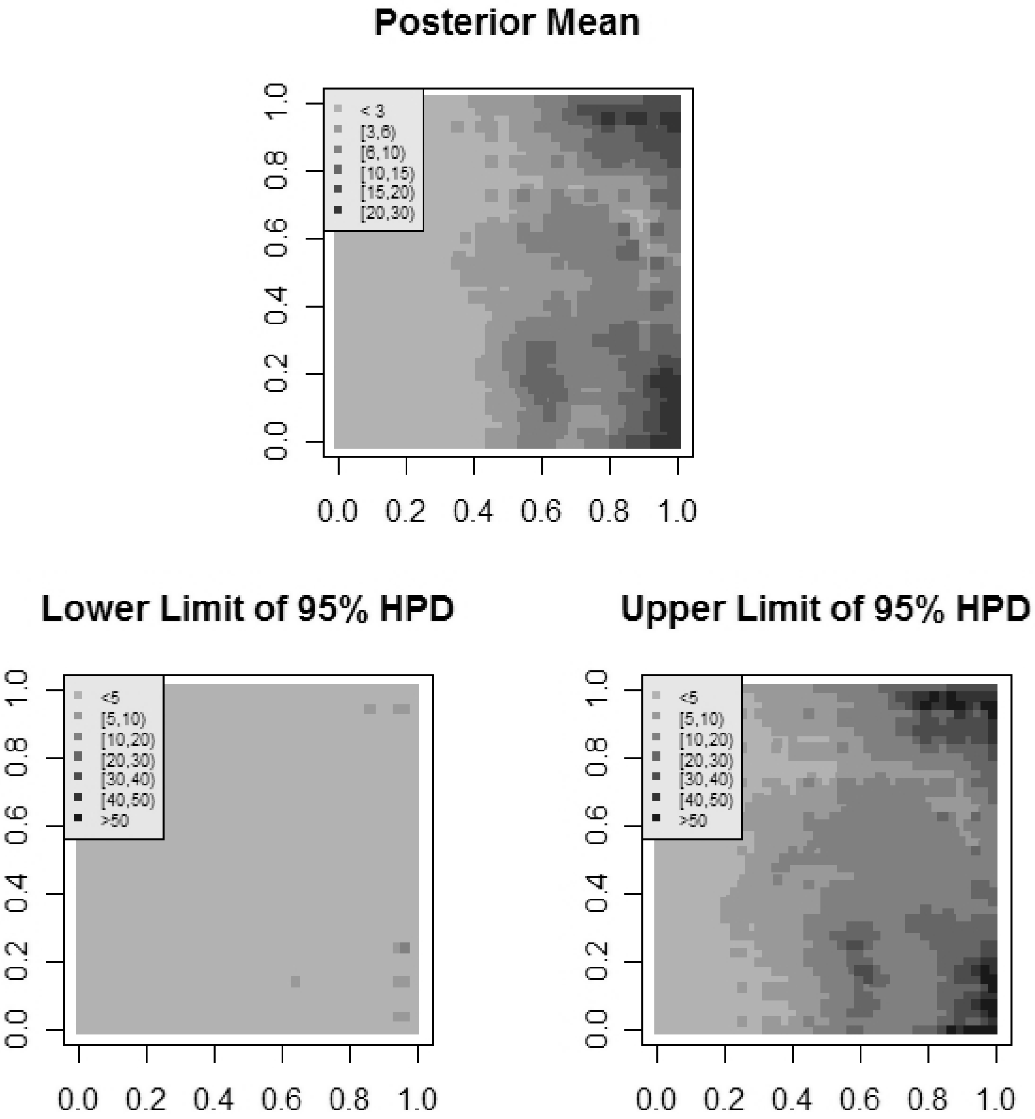


Fig. **5.16**: Posterior mean and 95% HPD for the semicontinuous response *Y* using the TSNC model.

5.6 Results Using the TSIBNC Model

In this subsection we simplify the covariance structure further by assuming independence between the two spatial stochastic processes, essentially modeling the processes of incidence and abundance separately. This is analogous to the approach of Ver Hoef and Jansen (2007) who assume Gaussian Markov random field models for the two processes on a lattice where we assume Gaussian process models for the two processes on a continuous spatial domain. For the TSIBNC model we employ the same covariance structures in Eq. **5.3**, simply setting **S** as a vector of independent normal random effects and an exponential covariance structure for **Z.** However, we remove the cross-covariance between the two variables and set $\boldsymbol{\rho}_{SZ}$ =0. We implemented MCMC using the algorithms described in Section 4.4. and additional details in Section 5.2.

Figure **5.17** shows posterior samples of the covariance and regression parameters, and the posterior distributions are summarized in Table **5.6**. The variance and regression parameters for the incidence part ($\sigma_{S,}$ $\alpha_0,$ α_1) are very similar between the TSIB (Figure **5.11**) and TSIBNC (Figure **5.17**) because the *S* random effects were assumed to be independent in both models. Likewise the posterior samples of regression parameters for abundance ($\boldsymbol{\beta}$) for these two models are also very similar. This is not surprising since the posterior sample values for ρ_{SZ}, the correlation between the two sets of random effects, was effectively zero in the reduced covariance model in TSIB. This means that the posterior distribution essentially already had two independent processes with separate covariances. The only difference is that the spatial dependence among the *Z* random effects is now better captured in the posterior distribution of this model. This is because the *Z* random effects are now just the "residual effects" in the abundance part of the data and are no longer affected by the lack of spatial dependence among the *S* random effects in the incidence part of the model. It is now apparent that when we impose correlation between the apparently independent process *S* and the spatially dependent *Z*, the result is a dilution in the correlation among the *Z*'s rather than being able to rein in the *S* random effects into a similar spatial structure.

Table **5.6**: Posterior estimates, Monte Carlo standard errors, and HPD for parameter estimates in the simulated data using the TSIBNC model.

Parameter	True Value	Posterior Mean (Standard Deviation)	Monte Carlo Standard Error	95% HPD
σ_S	1.0	6.04 (0.22)	0.009	(5.60, 6.46)
θ_Z	5.0	10.15 (2.25)	0.080	(6.53, 15.01)
α_0	2.0	7.36 (0.50)	0.028	(6.42, 8.37)
α_1	5.0	15.50 (1.00)	0.077	(13.41, 17.36)
β_0	1.0	1.67 (0.06)	0.002	(1.56, 1.79)
β_1	3.0	0.99 (0.10)	0.004	(0.78, 1.18)

Figure **5.18** shows the predicted surfaces for both sets of spatial random effects, as well as mean incidence $\left(E\left(U\right)\right)$ and mean positive$\left(E\left(V\right)\right)$ counts. By construction, the *S* random effects are uncorrelated, and this is reflected in the absence of any spatial dependence or clustering in the predicted surface. The map of mean incidence $\left(E\left(U\right)\right)$is then mostly an increasing function of distance from the origin of the X-axis.

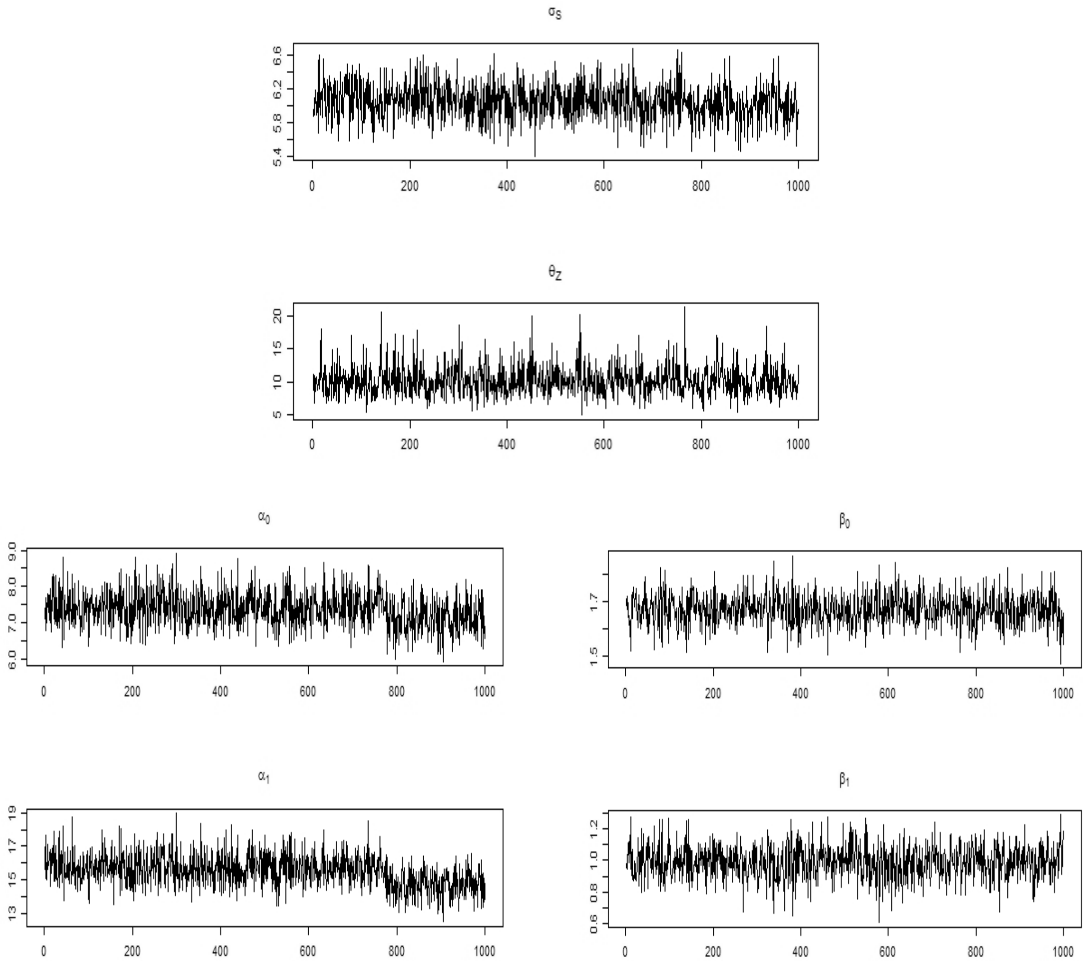


Fig. **5.17**: MCMC samples of covariance and regression parameters under the TSIBNC model.

Figure **5.19** maps the posterior mean and 95% HPD of predicted values of the semicontinuous variable *Y*, as we did in Figures **5.7** , **5.13** and **5.16** for the other covariance models. This and the earlier maps are consistent with the simulated values in Figure **5.2**, generally increasing along the x-axis with some areas of higher abundance that reflect clusters of higher Z random effects at the top right and bottom right corners.

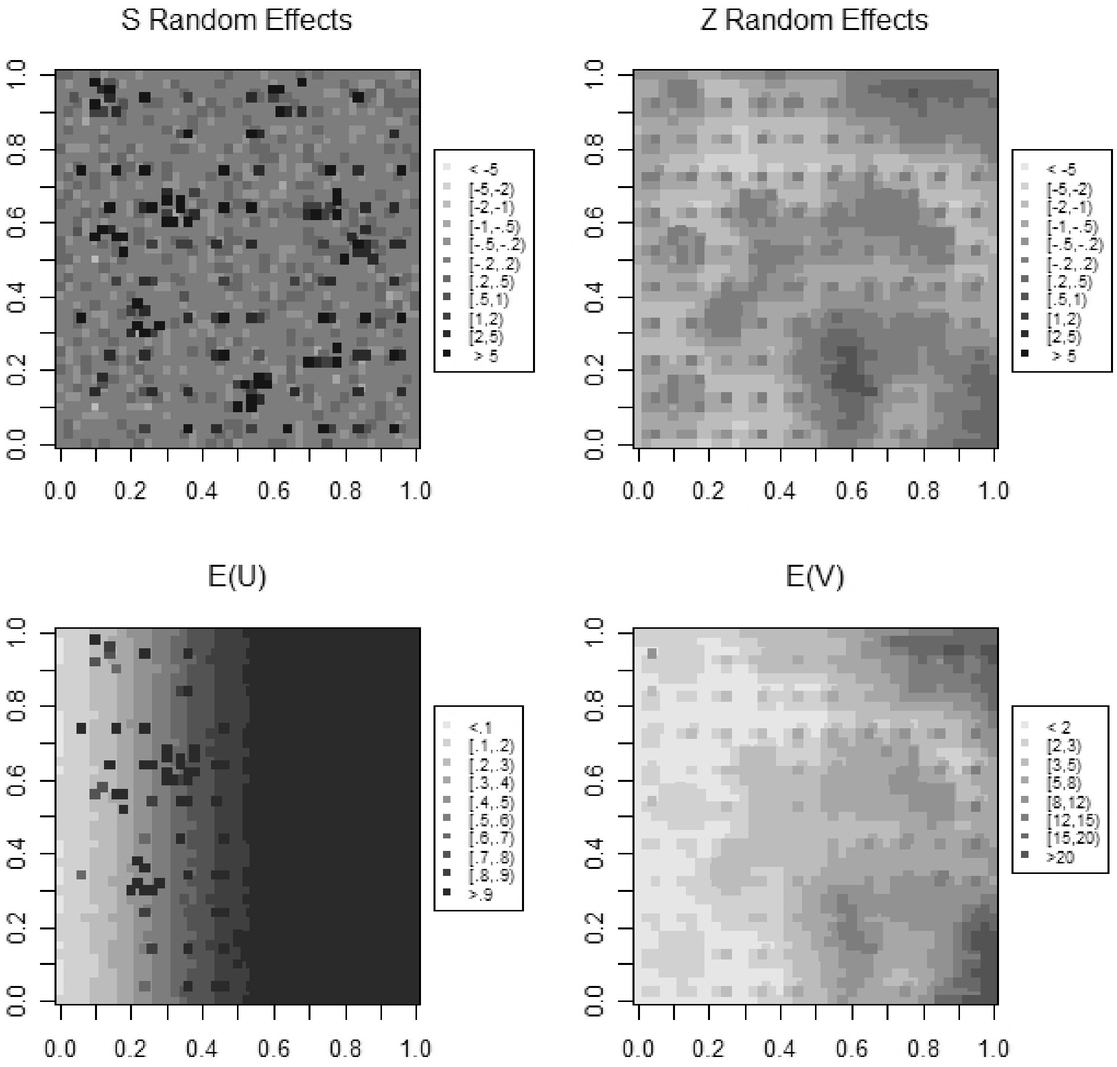


Fig. **5.18**: Predicted mean surface for simulated data using the TSIBNC model .

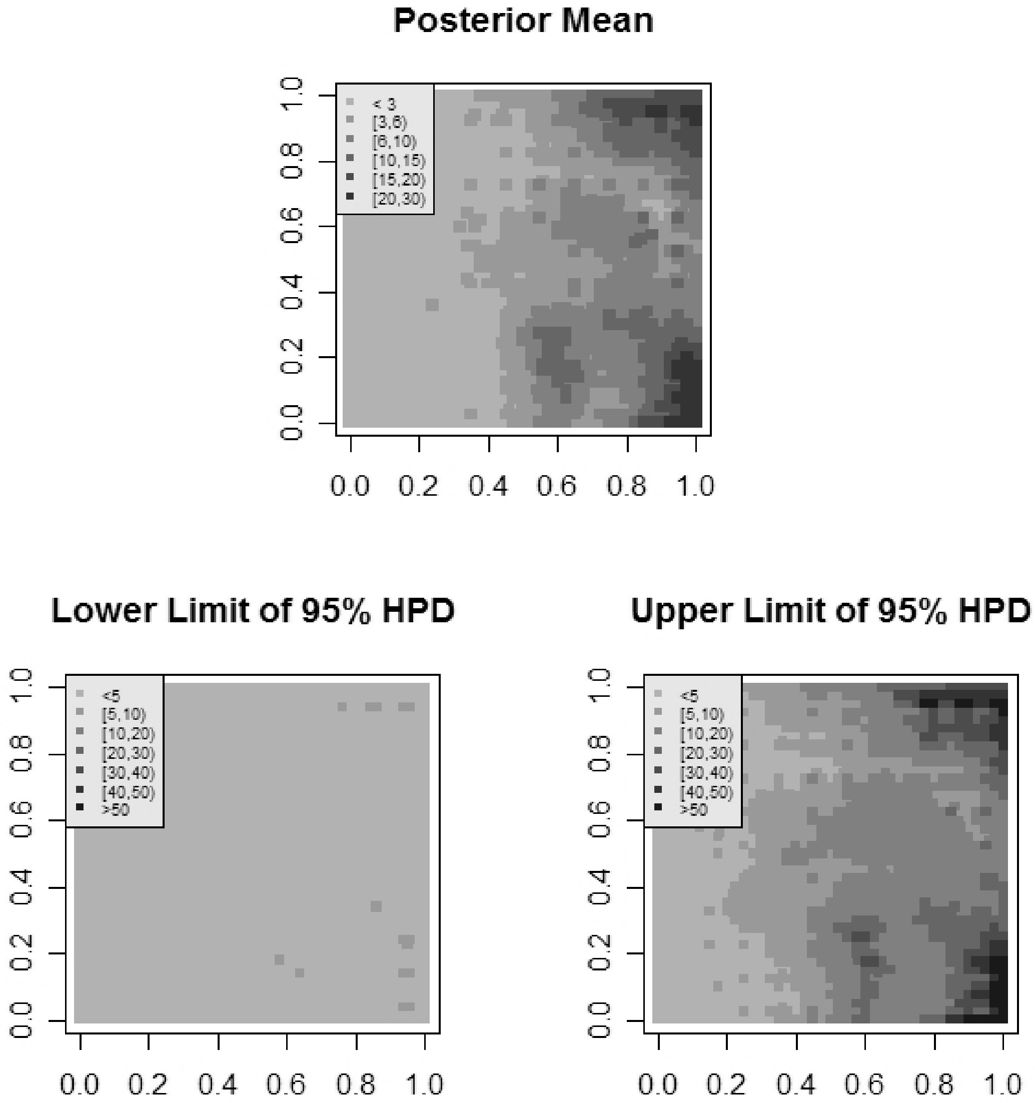


Fig. **5.19**: Posterior mean and 95% HPD for the semicontinuous response *Y* using the TSIBNC model.

5.7 Discussion

In this chapter we applied the two-part model presented in Chapter 3 on a simulated data set. We designed a sampling plan to estimate regression coefficients as well as covariance parameters by ensuring that there are sample locations scattered throughout the field to capture mean trend and, at the same time, enough closely-spaced locations to capture the behavior of the covariance function close to the origin. The allocation of these samples may be modified depending on whether the study is intended more to establish overall trend or spatial patterns.

We used a Bayesian approach implemented via MCMC to obtain samples of the regression coefficients (α and β) as well as spatial random effects (**S** and **Z**) for the incidence and abundance part of the model, respectively. We generated maps of prediction surfaces for the incidence (**U**) and abundance (**V**) parts of the model and combined these into a prediction map of the semicontinuous variable. We also obtained posterior samples of the covariance parameters under four covariance model specifications. TSF is the full two stage model as it has the same structure used to simulate the random errors (**S, Z**), TSIB uses a two stage model with independent random effects in the binary part, TSNC is the same as TSF but without the cross-correlation between the two sets of random errors (**S, Z**), and TSIBNC also assumes no correlation between the two sets of random effects in addition to independent binary random effects**.**

Table **5.7** summarizes the results from the four models. We employed different covariance models for the incidence part and its relation to the abundance part, with similar results for the parameters of the abundance part and some differences in the parameters for the incidence part across the four models.

The covariance model for the abundance part was similar across the four approaches and the resulting estimates for θ_Z, β_0, and β_1 are comparable across models. We found that the spatial dependence among **Z** random effects is captured in varying degrees in these models. In particular, the HPDs for θ_Z in the TSF, TSNC, and TSIBNC are comparable and narrower under these models compared to TSIB. In the TSIB model, allowing **S** random effects to vary independently while at the same time correlating them with **Z** appears to cause instability in capturing the spatial association among the **Z** random effects, as shown by the wider HPD of (7.49, 27.38) for θ_Z. When a correlation structure is imposed on the **S** (as in the TSF) or the cross correlation to the apparently unstructured **S** was removed (as in TSNC and TSIBNC), the posterior distribution more effectively assessed the spatial structure among the **Z** random effects.

Table **5.7**: Summary of HPD of MCMC samples for the different parameters under the four models for the simulated data.

Parameter	True Value	95% HPD[a]			
		TSF Model	TSIB Model	TSNC Model	TSIBNC Model
σ_S	1.0	--	(3.12, 6.88)	--	(5.60, 6.46)
θ_S	10.0	(76.8, 148.4)	--	(94.8, 148.4)	--
θ_Z	5.0	(6.30, 14.27)	(7.49, 27.38)	(6.34, 15.01)	(6.53, 15.01)
ρ_{SZ}	0.75	(-0.16, 0.21)	(-0.20, 0.16)	--	--
α_0	2.0	(1.95, 3.22)	(4.00, 8.65)	(1.94, 3.22)	(6.42, 8.37)
α_1	5.0	(4.69, 7.07)	(9.71, 20.27)	(4.71, 7.05)	(13.41, 17.36)
β_0	1.0	(1.56, 1.78)	(1.51, 1.74)	(1.56, 1.79)	(1.56, 1.79)
β_1	3.0	(0.78,1.18)	(0.84, 1.24)	(0.78, 1.19)	(0.78, 1.18)

[a] The HPDs for the parameters in each column are generated under different model assumptions and are not directly comparable. In particular, the regression parameters for the logit part of the model (α_0, α_1) should be interpreted conditionally on their respective random effects, which vary between the different models.

It was difficult for the distribution to capture **S** random effects because the binary outcome (presence vs. absence) provided very little information regarding its magnitude. However, some spatial structure in the **S** random effects can be gleaned by using the TSF or TSNC model because it incorporates a correlation among the **S,** as well as cross-correlation to a spatially dependent **Z** in the case of TSF. We also examined the performance of the two-stage model using less structured covariances, assuming independence among **S** random effects in the TSIB and TSIBNC models. Unfortunately, allowing the **S** random effects to vary independently resulted in considerable over-estimation of the regression coefficients for incidence (α_0, α_1). In modelling a binary random field by clipping a Gaussian random field, de Oliveira (2000) observed that "the inference about the binary map depends heavily on the correlation structure of the underlying Gaussian random field." By varying our covariance and cross-covariance models, we effectively changed the posterior distribution of the spatial random effects for the binary regression. We observe that when the correlation among the *S* was removed (as in the TSIB and TSIBNC models), marginal posterior means of the regression parameters were considerably different compared to those under the TSF and TSNC.

For the incidence part of the model, the marginal posterior means from both the TSF and TSNC models were closest to the true values and the 95% HPD for the sample α for these two models contain the true values used to generate the observations. However, this comes with the caveat that in the case of generalized linear mixed models, the regression parameters have a conditional rather than a marginal interpretation. In the case of these four covariance models, each model generates a different set of random effects, and the regression parameters are sampled conditional on each set of random effects. Therefore, the marginal posterior means are not necessarily comparable, and posterior means that are closest to the known parameters does not by itself make this model the best performer. Paraphrasing Diggle *et al.* (1998), "$E\left[U\left(x_i\right)\middle|S\left(x_i\right)\right]$ and $E\left[U\left(x_i\right)\right]$ differ in their structural dependence on the explanatory variables $\mathbf{d}\left(x_i\right)$, so the interpretation of α requires care. Only in the case where $U_i\left|S\left(x_i\right)\right.$ is Gaussian and the link function is the identity can α be treated as the regression parameter for the marginal regression function $E\left[U_i\right]$."

All four models produced similar prediction maps with features that are consistent with the simulated data. For probability of incidence $E\left(U\right)$ in Figures **5.6**, **5.12**, **5.15** and Figure **5.18**, the predictive maps generally followed a similar increasing trend from the left to the right side of the field, but the prediction for TSF and TSNC are smoother than those from models TSIB and TSIBNC. The difference is mainly due to smoothness of the random effects *S* associated with the incidence model, which are assumed to be independent in the TSIB and TSIBNC models, but have a dependence structure in TSF and TSNC. Although the posterior distributions for the TSF and TSNC do not capture the strong dependence that we used to simulate the data, they appear to have captured enough dependence to smooth the posterior values. On the other hand, when these random effects are allowed to vary independently in the TSIB and TSIBNC models, the random effects behaved more like "residuals" that compensated for presence or absence in each location by taking on large positive or large negative values, respectively.

Predicted abundance $E\left(V\right)$ maps were generally similar across the models because the marginal posterior distributions for the regression and covariance parameters are similar. And, because much of the patterns in the mean predicted values $E\left(Y\right)$ are derived from $E\left(V\right)$, not surprisingly the map of mean predicted values of Y are also similar across models. Figure **5.20** plots the posterior mean predicted value of the semicontinuous variable Y in each location (i.e., the values mapped in Figures **5.7**, **5.13**, **5.16**, and **5.19**) against the actual simulated value for that location (Figure **5.2**). The mean predicted values were consistent with the simulated values in most cases. The 95% HPD for the predicted values contained the actual observation 95% - 96% of the time. This predictive performance and HPD coverage is demonstrated in Figure **5.21** for a sample of values from the simulated data set. The actual simulated value of Y in 100 previously unobserved locations (i.e., not included in the n=400 used in the estimation) is plotted, along with posterior mean predicted Y values and 95% HPDs generated under each model. The values are ordered according to posterior mean predicted value, for ease of presentation. For all the models, the prediction HPDs provide good coverage of the simulated values.

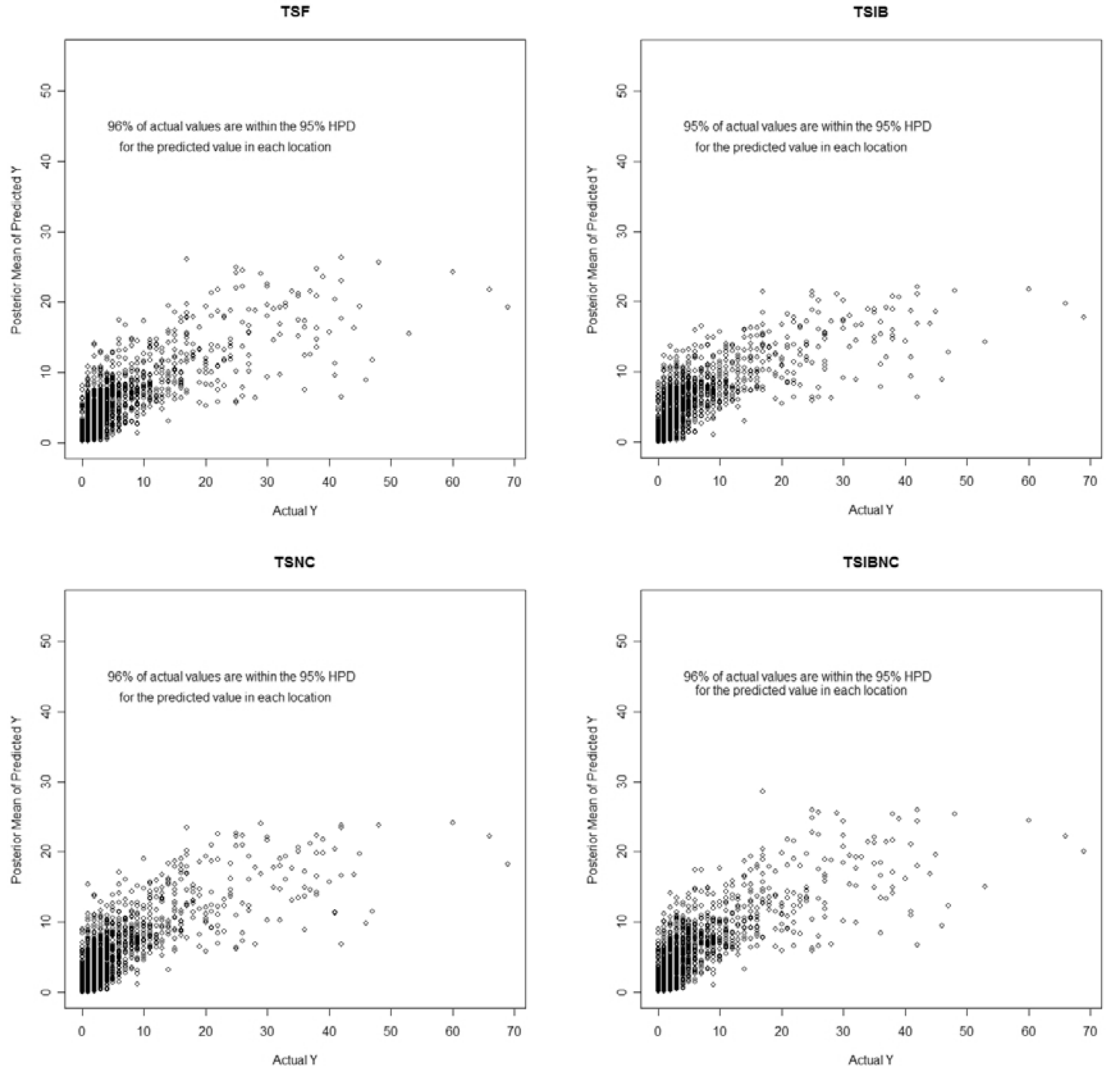
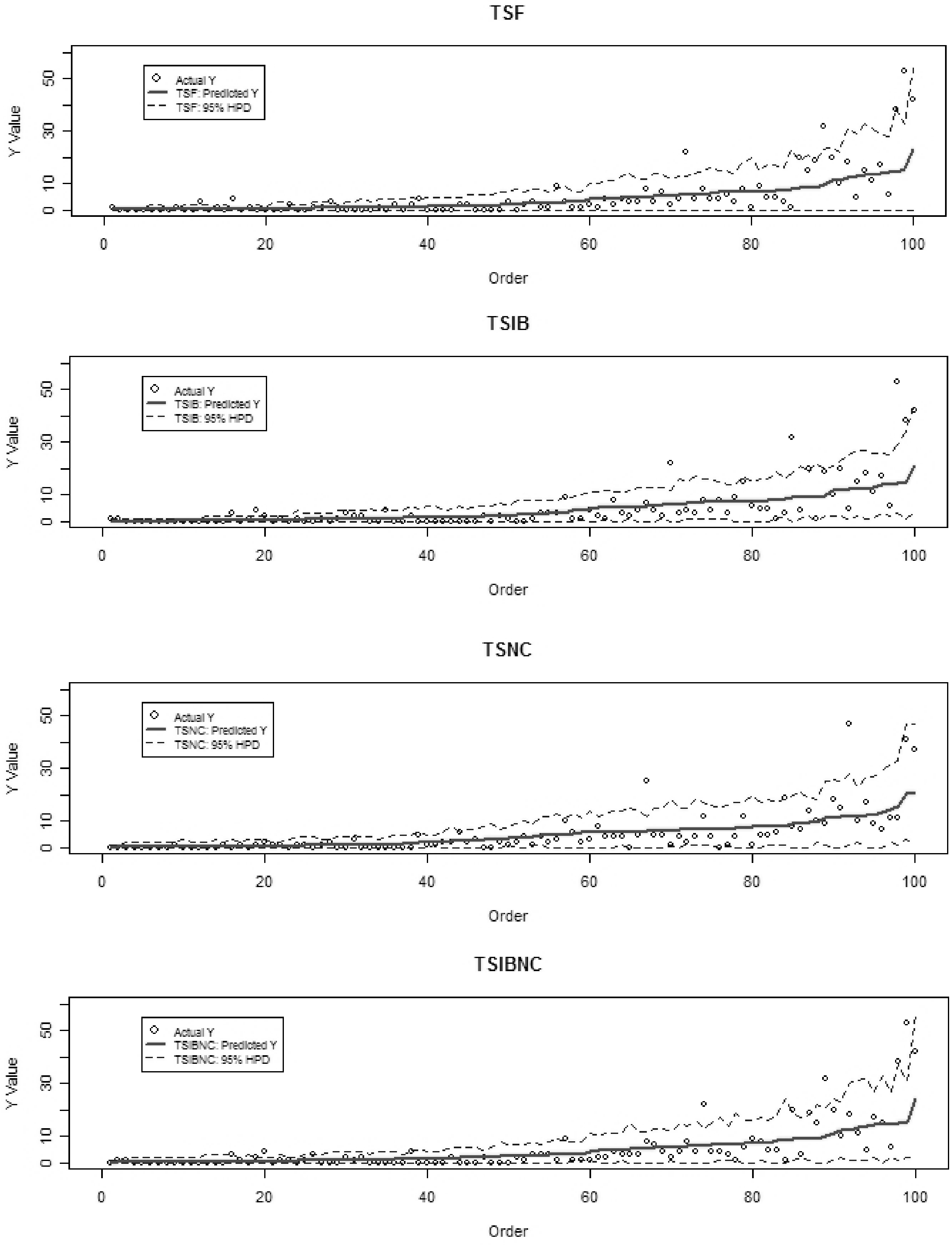


Fig. **5.20**: Scatterplot of actual simulated value of the semicontinuous variable *Y* against posterior mean predicted value under the four models.



TSF

TSIB

TSNC

TSIBNC

Fig. **5.21**: Posterior mean predicted value and 95% HPD for a sample of 100 unobserved values of *Y*.

The upper limit of HPDs missed the actual values in a small percentage (4-5% overall) of the locations, most notably when the actual value is much higher than the predicted value of *Y*, as shown in Figure **5.21** for the sample from the full data set. It is also helpful to see where these missed predictions are located. Figure **5.22** plots the upper limit of the 95% HPD for predicted value of *Y* (i.e., the same values mapped in Figures **5.7**, **5.13**, **5.16**, and **5.19**) only for the locations where the actual value (shown in its entirety on the bottom map) was higher than the upper limit of the corresponding HPD. It is evident that the predictions were missed mostly in the areas with very high observed counts, also called "hot spots". This is clearly a consequence of the covariance structure we have imposed on the random effects, which then tends to oversmooth the predicted random effect surface.

For each model, we computed a mean square error (MSE) of prediction for $E\left(U\right)$ by taking $MSE_U = \frac{1}{n_{\mathbf{x}*}} \sum_{x\in\mathbf{x}*} \left[E_{\hat{\pi}}\left(U\left(x\right) \mid Data\right) - E\left(U\left(x\right) \middle| \boldsymbol{\alpha}, S\left(x\right)\right)\right]^2$, where $E_{\hat{\pi}}\left[U\left(x\right) \mid Data\right]$ is the posterior mean of the probability of incidence given the incidence regression parameters and *S* random effect which we generated over all unobserved locations $x \in \mathbf{x}*$, as previously described in Section **5.2**. We computed MSE_V and MSE_Y similarly for $E\left(V\right)$and $E\left(Y\right)$.

These MSEs are summarized in Table **5.8**. The overall predictive performances of the models are comparable, although TSIB appears to be numerically inferior to TSF, TSNC and TSIBNC particularly with respect to predicting $E\left(V\right)$ and $E\left(Y\right)$.

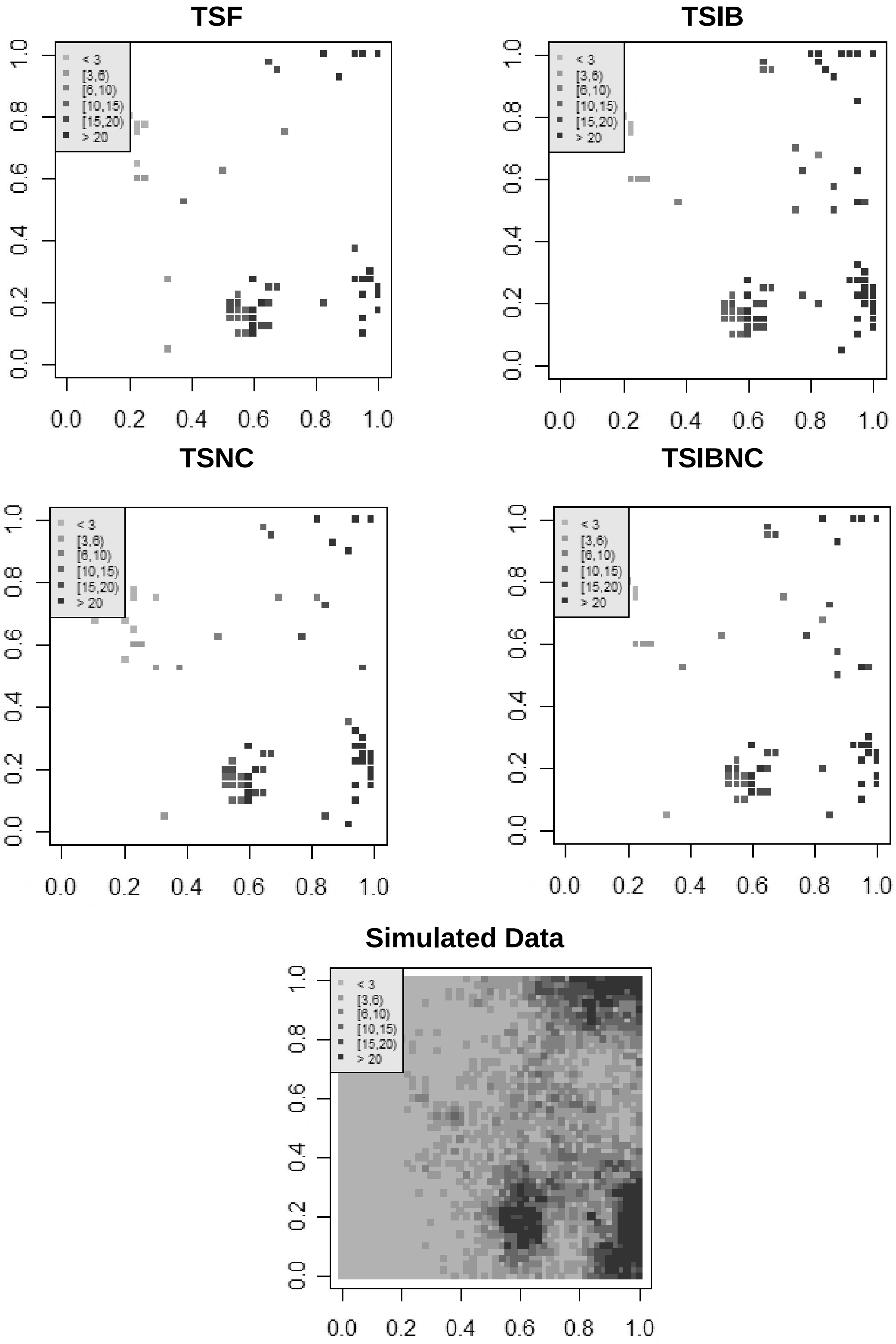


Fig. **5.22**: Locations where the upper limit of 95% HPD for predicted *Y* was lower than the actual simulated value.

Table **5.8**: Estimated mean square errors of prediction under each model for $E(U), E(V)$ and $E(Y)$.

	TSF Model	TSIB Model	TSNC Model	TSIBNC Model
MSE$_U$	0.018	0.012	0.012	0.013
MSE$_V$	22.42	25.58	22.47	22.64
MSE$_Y$	26.99	29.38	24.49	26.60

Whether the objective is estimation or prediction, we recommend the TSF or TSNC models. In this simulated data set, spatial dependence was only weakly identified by these models, particularly in the incidence part. However, this appears to have been enough to be able to capture the regression parameters for the incidence part and to produce predicted incidence maps that are reasonably smooth, consistent with our a priori assumption of spatial dependence. Both TSIBNC and TSIB produce posterior distributions for the parameters of the positive part of the model comparable to TSF and TSNC, but the latter models retain the mechanism to capture spatial dependence in both parts of the model. Cross-correlation may be kept or removed depending on prior information or as a tool to perform sensitivity analysis, exploring any changes in the other marginal posteriors under different levels of cross-correlation.

It can be said that TSIBNC can perform just as well as TSF or TSNC because it is more straightforward to assess spatial random effects solely from the abundance part in TSIBNC. Assuming independence between the two sets of random effects also makes the model more parsimonious and has the added advantage of working with much smaller covariance matrices, because we do not have to model **S** and **Z** jointly using their combined covariance matrices with cross-covariance. However, in this research we are motivated by the presence of excessive zeros in spatial data, and modeling the incidence process carefully is central to our goals since the zero observations from the data arise from the incidence process. This means that the incidence part should perform well on its own because all the zero observations in the data arise from this part of the model. Based on our study of this simulated data set, we find that the TSF and TSNC models enable a careful modeling of both the incidence and abundance parts, and therefore these are the models we would always use first before the simpler models.

We end this chapter with additional comments about spatial random effects and the consequences of model choice. In our simulated data set the Z random effects were negatively correlated with the lone covariate $d(x)$. This association was clearly just a chance occurrence; another realization may or may not have the same associations, or have different ones. However, in exploring the posterior distributions, we condition on *this* single realization, and therefore these associations, whether by chance or by design, are real and can greatly affect our findings; in this case the regression coefficients for the abundance model differed substantially from the parameters we used to generate the data.

In sampling the posterior distributions we assume that the Zs are random, but we are in fact trying to estimate the magnitude of this particular realization of the "random effect", using our belief about the process that generated it (in our case, a Gaussian stochastic process with a specified covariance). In this context, Reich *et al*. (2006) suggest that these are implicit fixed effects, and the posterior mean and variance of the regression coefficients can be affected when there is collinearity between the known covariates in the model and the spatial random effects.

In real data sets, it is possible that random effects become surrogates for other covariates that are not included in the model. The potential exists for confounding between the fixed effects in the model and the omitted variables that affect the posterior distributions through random effects. In our two-part model, it is particularly important to consider covariates that may be overlooked because they only affect the abundance part of the model.

Chapter 6

An Application to Ecology

In this chapter we revisit the entomological study in which different life stages of Colorado potato beetle were counted weekly at a resolution of one meter-row. We apply the same two-part approach as in the previous chapter, using the four models TSF, TSIB, TSNC, and TSIBNC.

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 **6.1** plots categorized densities observed in the sampled locations. The 296 observations have 144 zeros and 152 positive counts (Blom and Fleischer, 2001; Blom *et al*., 2002).

Each observation was transformed into a two-part response (U_i, V_i) and the same model as the first case study was fitted: $U_i|S(x_i) \sim$ Bernoulli(A_i) and $V_i|Z(x_i) \sim$ Truncated Poisson(B_i). 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 Eq. **5.1**.

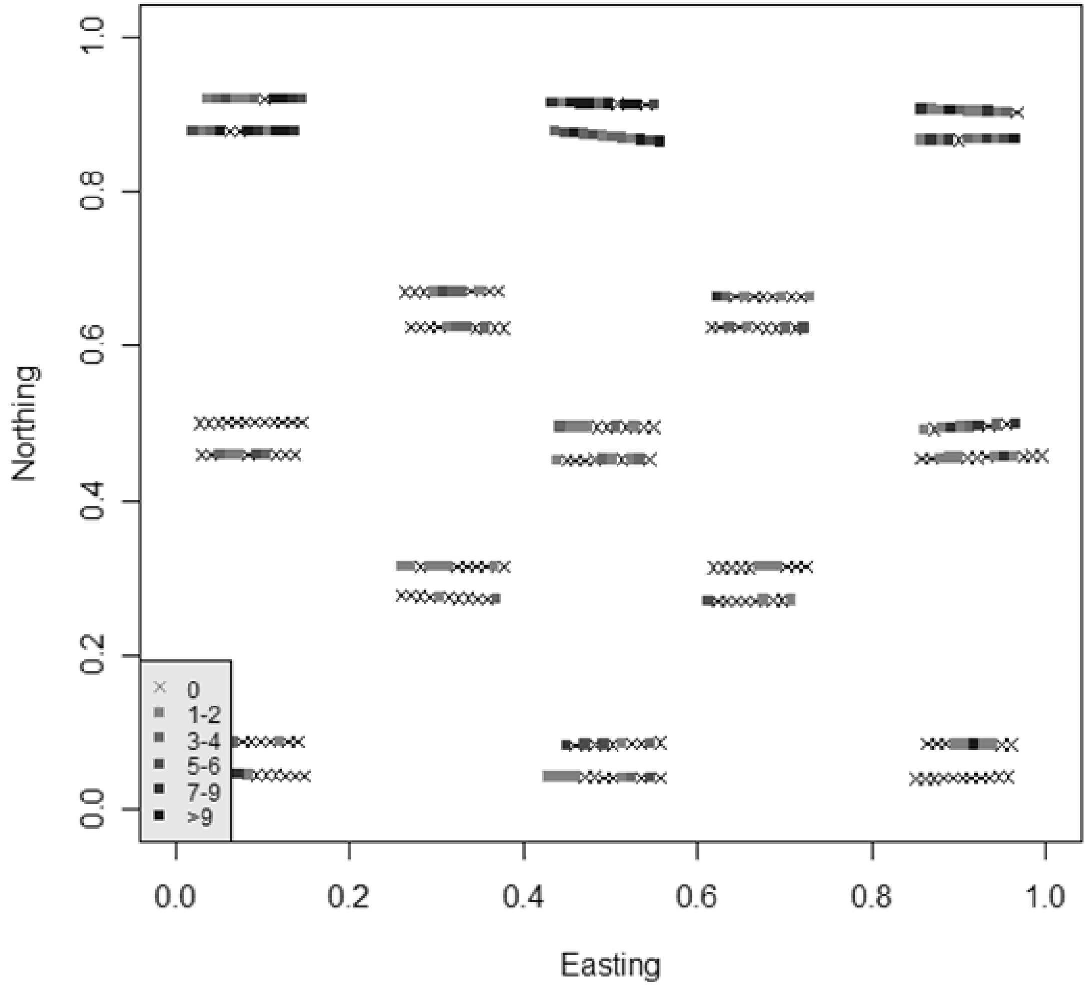


Fig. **6.1**: Observed counts of CPB large larvae at Week 8

6.1 Results from the TSF Model

We implemented MCMC using the algorithms described in Section **4.4** on the CPB data set, using the same independent priors as in Section **5.3** In generating posterior samples of **S** and **Z** following the LH algorithm described in **4.4.4**, we used truncation constant H = 50 for $\nabla(\gamma)^{trunc}$ and variance scale parameter h=0.40.

Figure **6.2** plots samples from the posterior distributions for the covariance and regression parameters at every 100th iteration after a sufficient burn-in period.

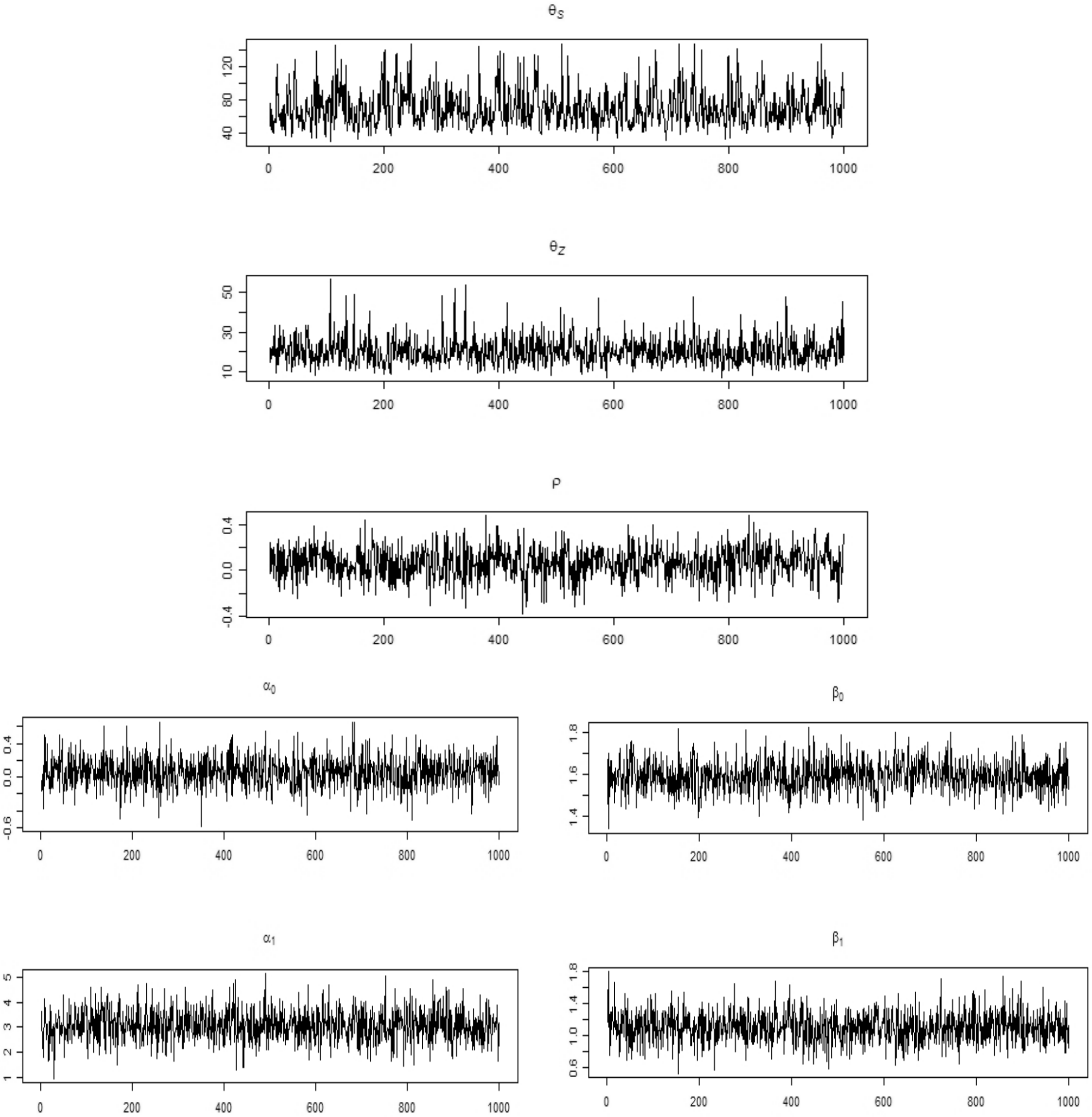


Fig. **6.2**: MCMC samples of all parameters for the CPB data set under the TSF model.

The marginal posterior distributions for the covariance and regression parameters were reasonably symmetric (Figure **6.3**), except for some skewness in θ_S and θ_Z.

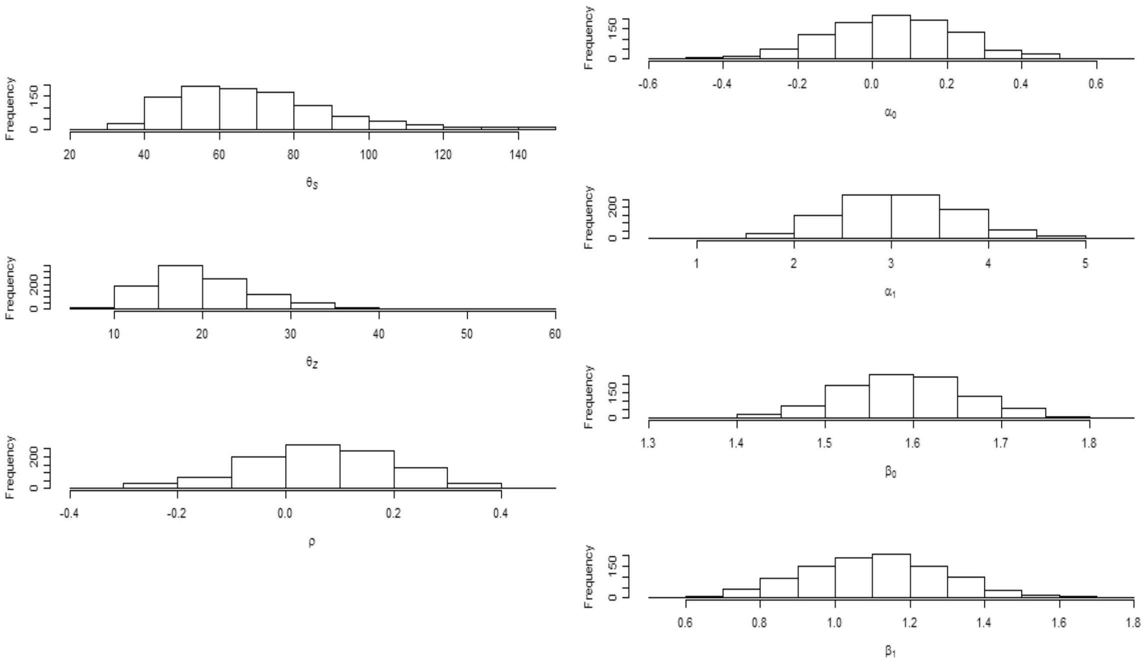


Fig. **6.3**: Distribution of MCMC samples of covariance and regression parameters for the CPB data set using the TSF model

Table **6.1** summarizes the results from the MCMC sampling procedure under the TSF model. The posterior mean for the θ_S samples is 70.81, showing weak spatial correlation. For instance, at the smallest observed distance of 0.88 m, two *S* signals are correlated by about 0.48, but at a distance of 2 m, the correlation drops to 0.2, and to about 0.08 at 3 m. The *Z* signals exhibit stronger spatial correlation, with a mean of 20.27 among the MCMC samples of θ_Z. At this value of θ_Z, at the smallest observed distance of 0.88 m, two *Z* signals are correlated by about 0.8; at a distance of 2 m, the correlation drops to 0.6, and then to about 0.3 among locations that are 5 m apart. The *S* and *Z* signals appear to be uncorrelated since the 95% highest posterior density (HPD) interval for the correlation is (-0.20, 0.34), which includes 0.

Table **6.1**: Posterior estimates, Monte Carlo standard errors, and HPD for parameter estimates from the CPB data set under the TSF model.

Parameter	Posterior Mean (Standard Deviation)	Monte Carlo Standard Error	95% HPD
θ_S	70.81 (22.49)	1.024	(32.48, 114.03)
θ_Z	20.27 (6.67)	0.191	(10.00, 33.56)
ρ_{SZ}	0.07 (0.14)	0.006	(-0.20, 0.34)
α_0	0.06 (0.18)	0.005	(-0.31, 0.41)
α_1	3.07 (0.64)	0.017	(1.81, 4.29)
β_0	1.59 (0.07)	0.003	(1.44, 1.72)
β_1	1.11 (0.19)	0.006	(0.71, 1.46)

Figure **6.4** maps the posterior mean of the S and Z samples, as well as mean predicted incidence $(E(U))$ and positive counts$(E(V))$. As in the results from the simulated data set, the *S* random effects appeared to be independent except for some localized areas of large positive random effects. The samples of *Z* random effects exhibited some spatial dependence shown as lighter or darker areas. Consequently, the map of mean predicted incidence $(E(U))$ shows a generally increasing mean trend along the y-axis, and some localized variation where positive counts were observed in the sampled locations. In these spots, the random effects are positive. The mean predicted positive counts $(E(V))$ shows a similar increasing trend along the y-axis as well locally higher or lower means that are consistent with the patches of lighter (negative) or darker (positive) *Z* random effects.

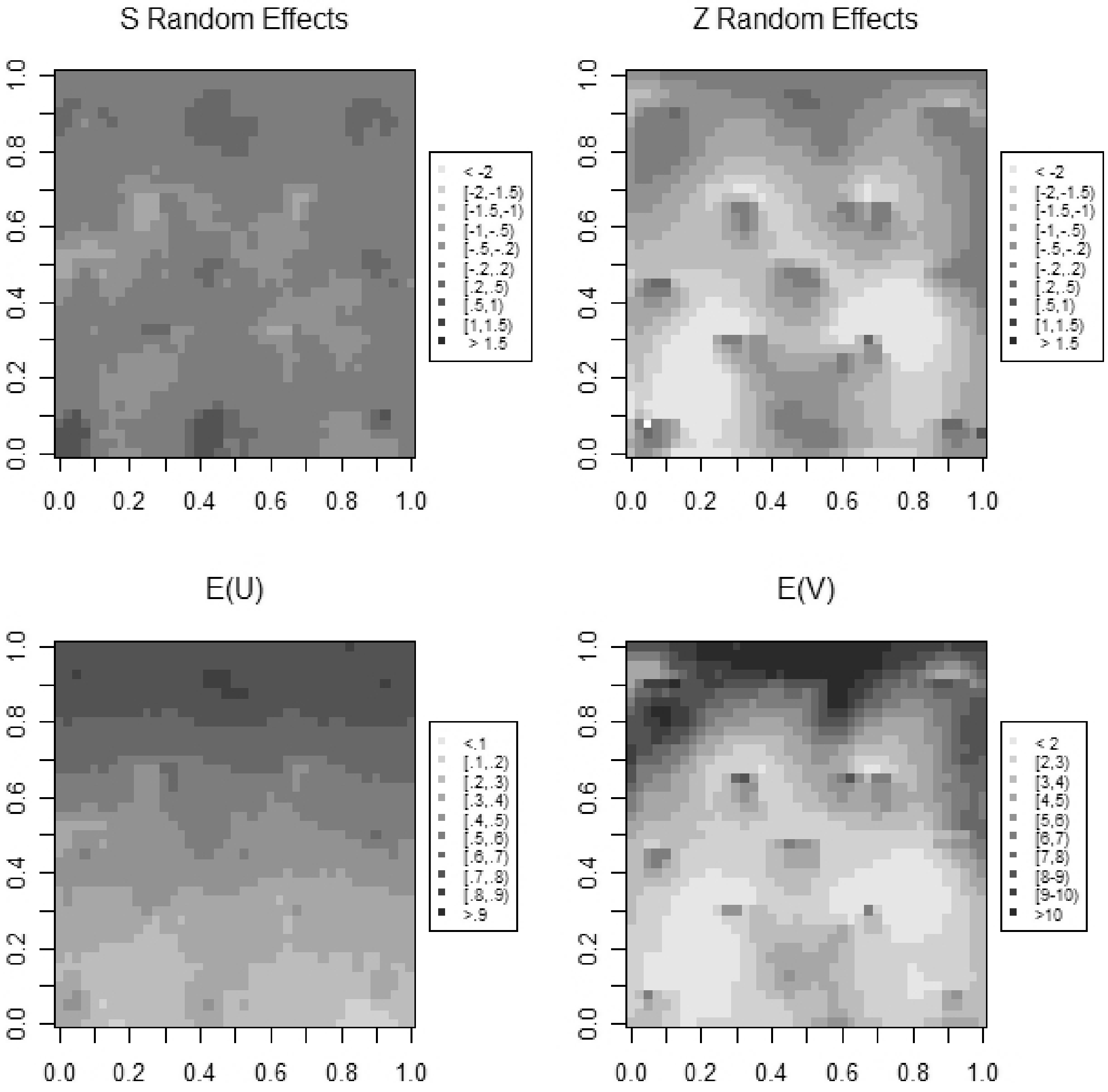


Fig. **6.4**: Predicted mean surface under the TSF model.

The positive mean of the posterior samples for both slopes (α_l, β_l) 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 3.07 (from Table **6.1**) as our estimate of α_l, the odds of finding at least one large larva increases from 1 in the middle of the field to about $e^{(0.5)(3.07)}=4.6$ to the north end of the field.

It is possible that the spatial processes that affect incidence (the U's) may be different from those that drive abundance (the V's). 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 abundance are really two different processes with different covariance structures. This can explain the apparent lack of correlation between the S and Z random effects, even in the same location. However, based on our finding in Chapter 5 regarding the difficulty in capturing S random effects and their correlation with Z, it is possible that the spatial random effects for incidence and abundance of large larvae are correlated, although not detected in the model.

Finally, we generated a realization U and V, and hence the semicontinuous variable Y, where $Y=0$ when U=0 and $Y=V$ when U=1. The posterior means of the Y and the 95% HPD for the predicted Y are mapped in Figure **6.5**. There is a clear increasing trend in the posterior mean as we move closer to the north edge of the field, although localized areas of higher or lower expected values are still evident. Areas with higher means have higher variability. This is not surprising since the positive component of the two-part model is the truncated Poisson, which is characterized by an increase in variance with increasing mean.

In addition to the mean, we can also map other functionals of interest. In this case, we generated the 95% HPD of the predicted values of *Y* and map them in Figure **6.5** as well. This can be used to identify possible "hot spots" or areas that may need control measures. We can also generate maps that show locations where predicted values exceed a known threshold.

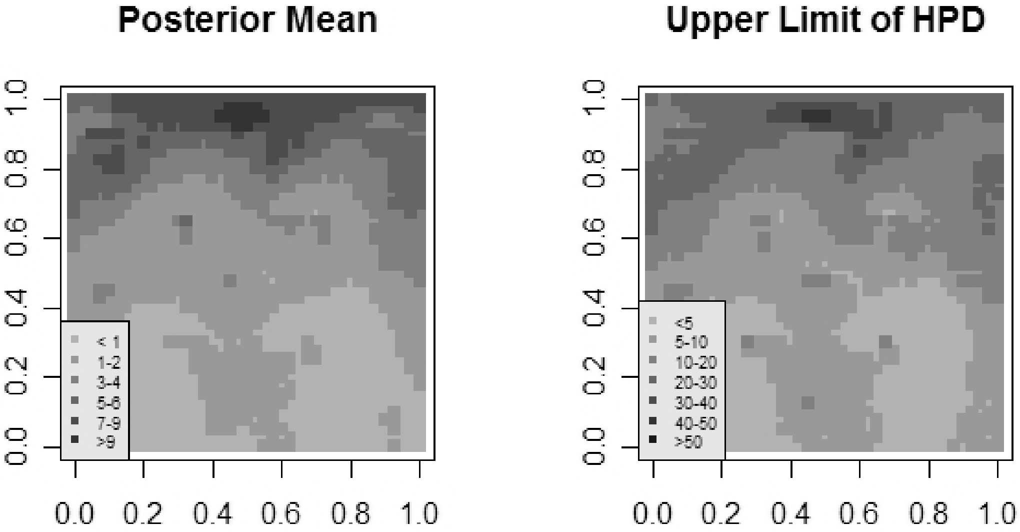


Fig. **6.5**: Posterior mean and upper limit of the 95% HPD for predicted CPB larvae per meter row in the observed field under the TSF model. The lower limit of the 95% HPD is 0 for all locations and is not mapped here.

6.2 Results from the TSIB Model

In this section we apply the two-part model on the same CPB data set, but this time model the covariance among random effects using the TSIB model, where we assume independence among the random effects in the binary part, as we did in Section **5.4** for the simulated data set. We employed the same priors as in the previous section. The ranges for the proper uniform priors for the covariance parameters (σ_S, ρ_{SZ}) are (0, 10] × [-1, 1]. We used improper flat priors for the regression parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Figure 6.6 shows the samples of covariance and regression parameters from the posterior distribution of the TSIB model.

Table 6.2 summarizes the posterior samples under the TSIB model. The posterior means for the covariance parameters that were retained in the model (θ_Z and ρ_{SZ}) were similar to those in the full covariance model (TSF), although the posterior mean and HPD for θ_Z was slightly higher in the TSIB model. This is the same trend seen in the simulated data set (Section **5.4**), where the correlation among the *Z* becomes less evident when the model includes a mechanism to correlate them to a spatially unstructured process (*S*). The regression parameters were also similar, except for an increase in the magnitude of α_I. In this model the posterior mean for α_I is 6.86 compared to 1.90 in the TSF model, showing a steeper mean gradient for incidence in the TSIB. However, these coefficients should be interpreted conditionally on the random effects in each model.

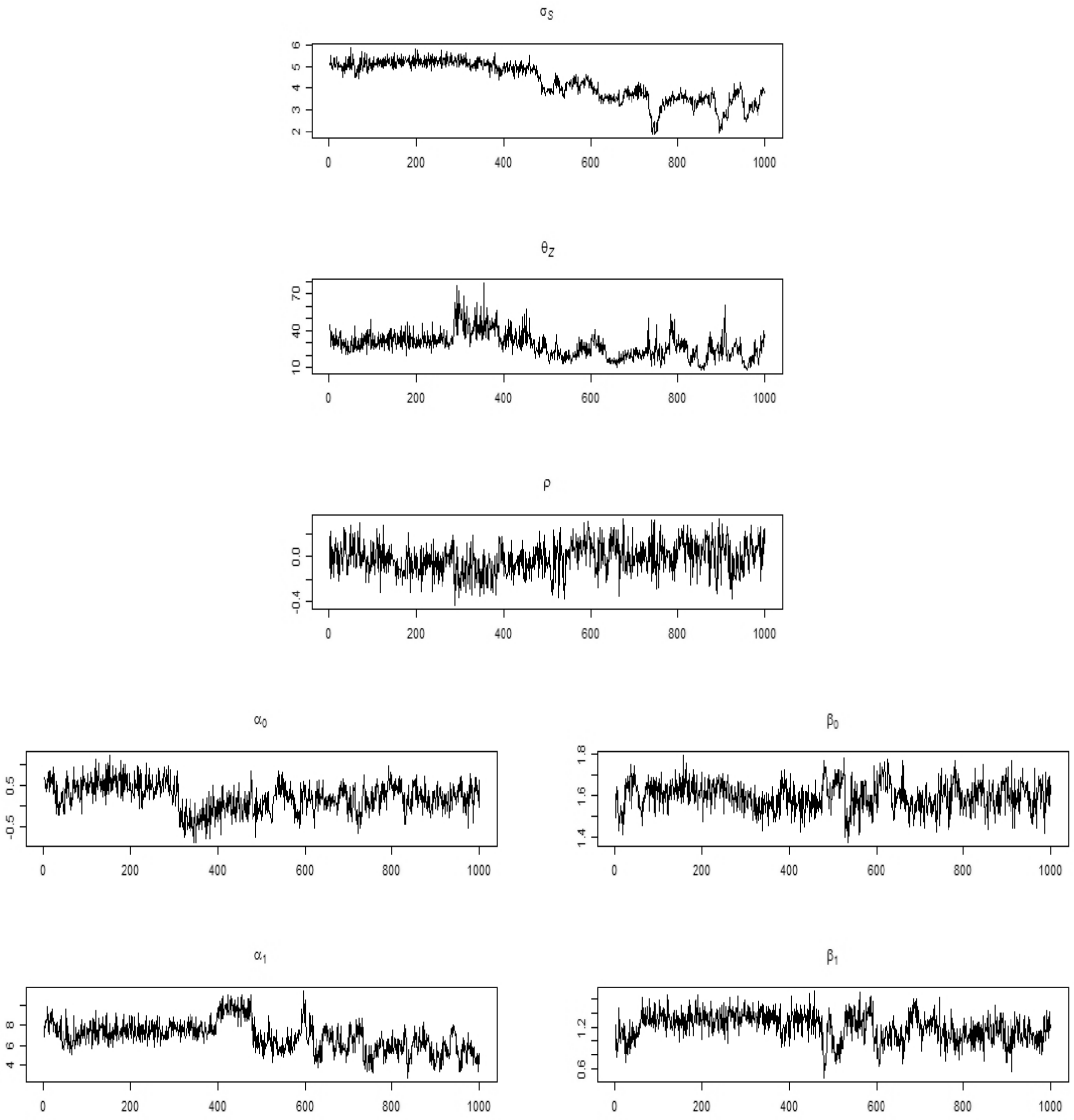


Fig. **6.6**: MCMC samples of all parameters for the CPB data set under the TSIB model.

Table **6.2**: Posterior estimates, Monte Carlo standard errors, and HPD for parameter estimates from the CPB data set under the TSIB model.

Parameter	Posterior Mean (Standard Deviation)	Monte Carlo Standard Error	95% HPD
σ_S	4.30 (0.89)	0.149	(2.61, 5.55)
θ_Z	28.47 (10.24)	1.398	(10.00, 48.09)
ρ_{SZ}	-0.01 (0.14)	0.010	(-0.27, 0.26)
α_0	0.22 (0.36)	0.045	(-0.47, 0.91)
α_1	6.86 (1.50)	0.200	(4.05, 9.99)
β_0	1.60 (0.07)	0.006	(1.45, 1.73)
β_1	1.20 (0.20)	0.023	(0.80, 1.57)

Figure **6.7** maps the posterior mean of the *S* and *Z* samples, as well as mean expected incidence $\left(E(U)\right)$ and positive counts$\left(E(V)\right)$. With this covariance function, the *S* random effects are assumed to be independent, and this is reflected in the absence of spatial persistence in the map of *S* random effects. In contrast, the posterior mean of the samples of Z random effects show spatial patterns similar to those in Figure **6.4** . The maps of expected incidence $\left(E(U)\right)$ and abundance $\left(E(V)\right)$ are generally increasing functions of proximity to the north edge of the field, as observed earlier. There are some localized areas of higher or lower abundance due to the spatial patterns seen among the Z random effects.

Maps of the posterior mean and 95% HPD of the semicontinuous response are presented in Figure **6.8**. The patterns seen here are similar to those in Figure **6.5** (corresponding to the TSF model). There is a clear increasing trend in the expected mean towards the north edge of the field, and an increase in variability as well.

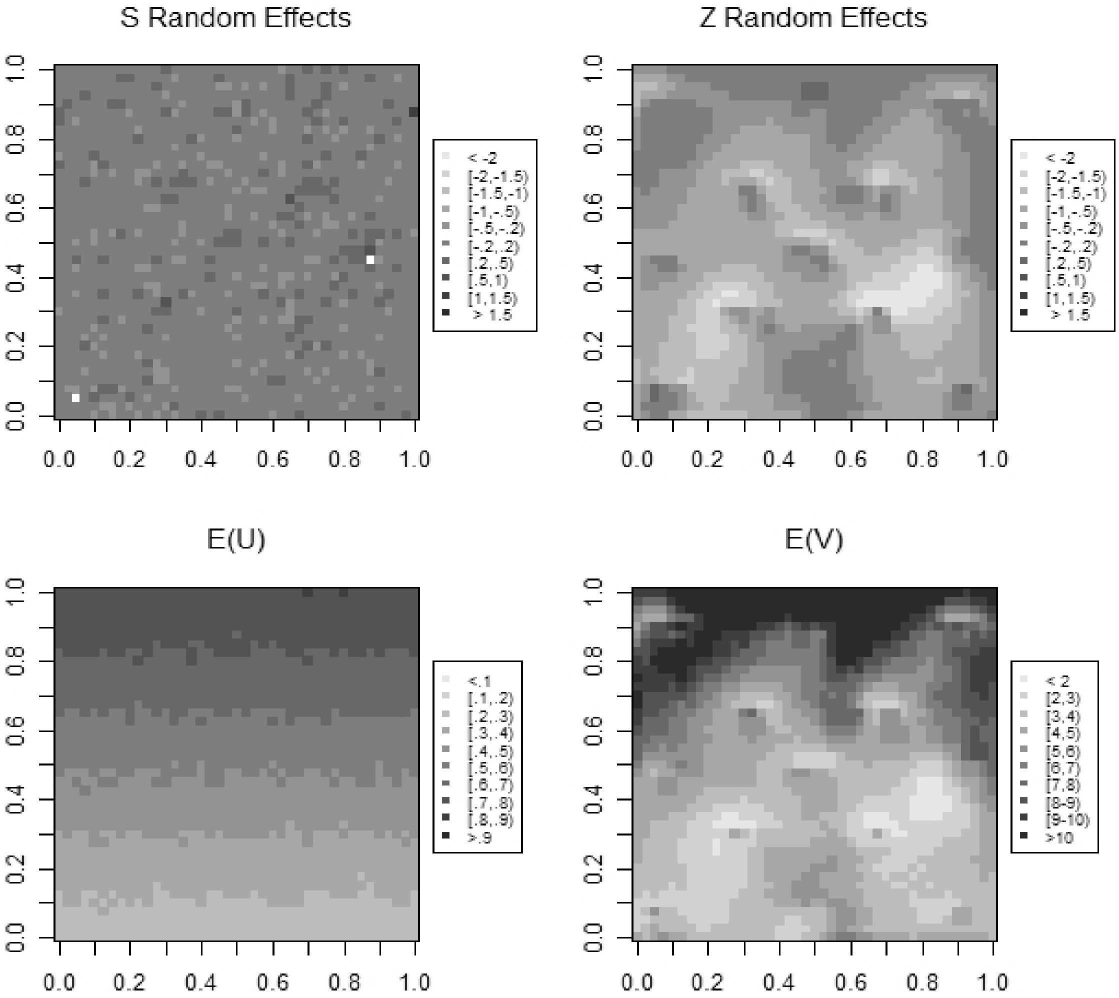


Fig. **6.7**: Predicted mean surface under the TSIB model.

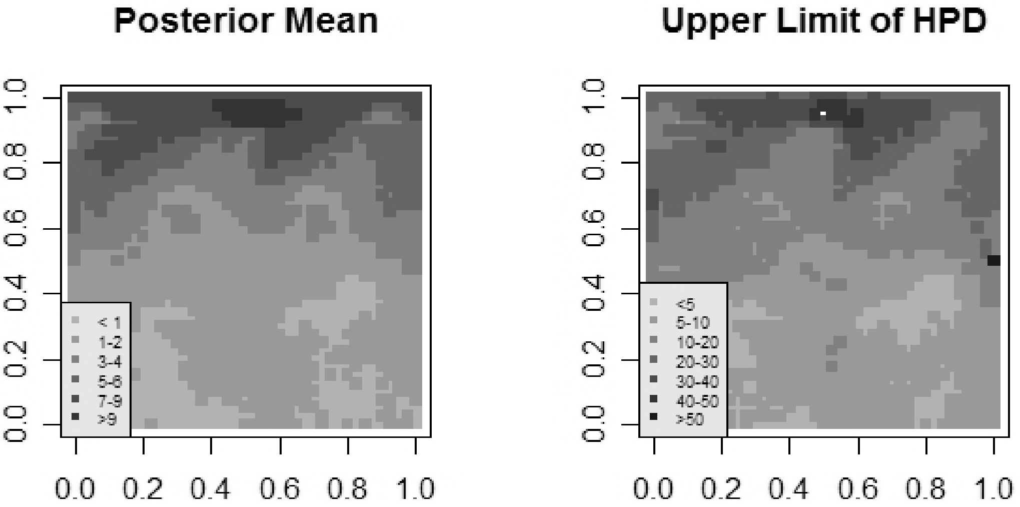


Fig. **6.8**: Posterior mean and upper limit of 95% HPD for predicted CPB larvae per meter row in the observed field under the TSIB model. The lower limit of the 95% HPD is 0 for all locations and is not mapped here.

6.3 Results from the TSNC model

The TSNC model assumes spatially dependent covariance functions for each set of random effects but removes cross-correlation by setting $\rho_{SZ} = 0$. We used the same priors as the TSF model in the MCMC implementation.

Figure 6.9 plots the posterior samples for all parameters and Table 6.3 summarizes the sample values from the posterior distribution of the TSNC model. The posterior distributions of the parameters for the abundance part appear to be unchanged relative to those in the TSF model in Section **6.1**. As in the application of TSNC to the simulated data, the correlation parameter ρ_{SZ} was essentially 0 in the TSF model, so the abundance part of the posterior model was being sampled independently from the incidence part. We note that the posterior mean and HPD for covariance parameter θ_S is slightly lower in TSF [mean=69.14, HPD= (33.72, 118.39)] than in this TSNC model [mean=85.74, HPD= (43.78, 148.41)], which means that the spatial dependence among *S* random effects was slightly more pronounced when cross-correlation is allowed, such as in TSF.

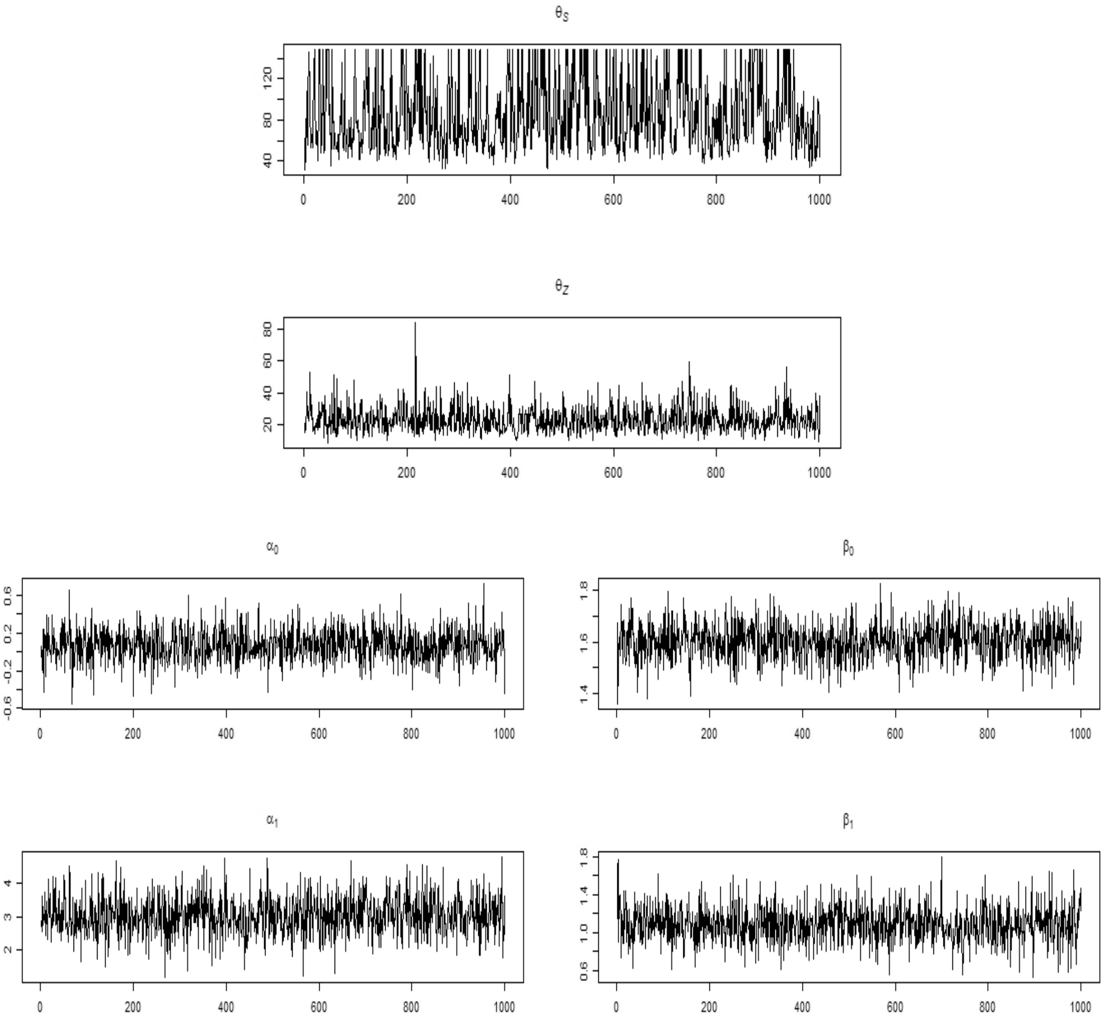


Fig. **6.9**: MCMC samples of all parameters for the CPB data set under the TSNC model.

Table **6.3**: Posterior estimates, Monte Carlo standard errors, and HPD for parameter estimates from the CPB data set under the TSNC model.

Parameter	Posterior Mean (Standard Deviation)	Monte Carlo Standard Error	95% HPD
θ_S	85.74 (34.94)	1.901	(43.78, 148.41)
θ_Z	23.44 (8.18)	0.321	(9.84, 39.27)
α_0	0.07 (0.18)	0.006	(-0.24, 0.42)
α_1	3.03 (0.60)	0.022	(1.94, 4.23)
β_0	1.60 (0.07)	0.002	(1.47, 1.75)
β_1	1.09 (0.20)	0.006	(0.73, 1.50)

Figure **6.10** maps the predicted surfaces for both sets of spatial random effects, as well as mean incidence $\left(E\left(U\right)\right)$ and mean positive$\left(E\left(V\right)\right)$ counts, and Figure **6.11** maps the posterior mean and upper limit of the HPD of the semicontinuous variable *Y*. These maps are similar to the corresponding maps in the TSF model, although a visual comparison appears to indicate that even less spatial dependence in the *S* random effects in the TSNC model compared to TSF.

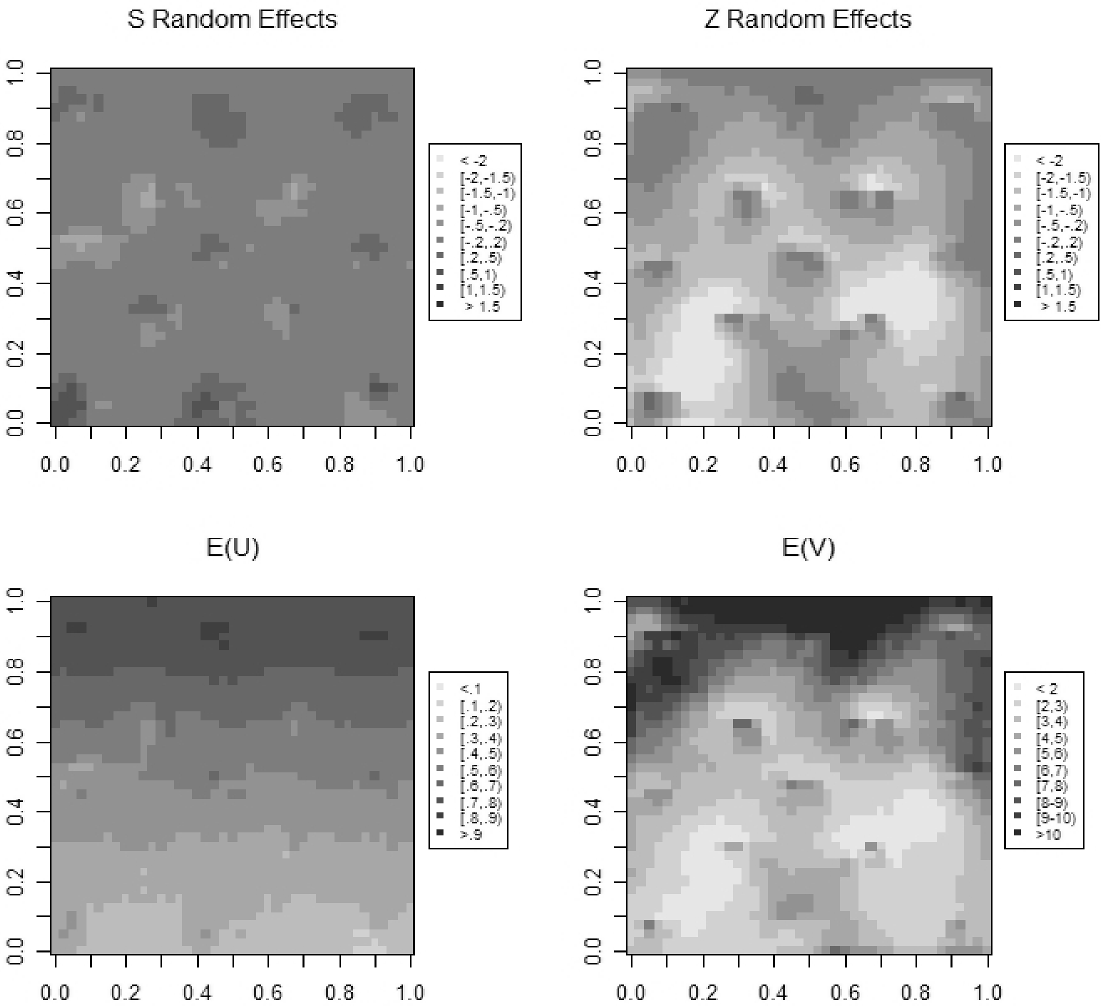


Fig. **6.10**: Predicted mean surface under the TSNC model.

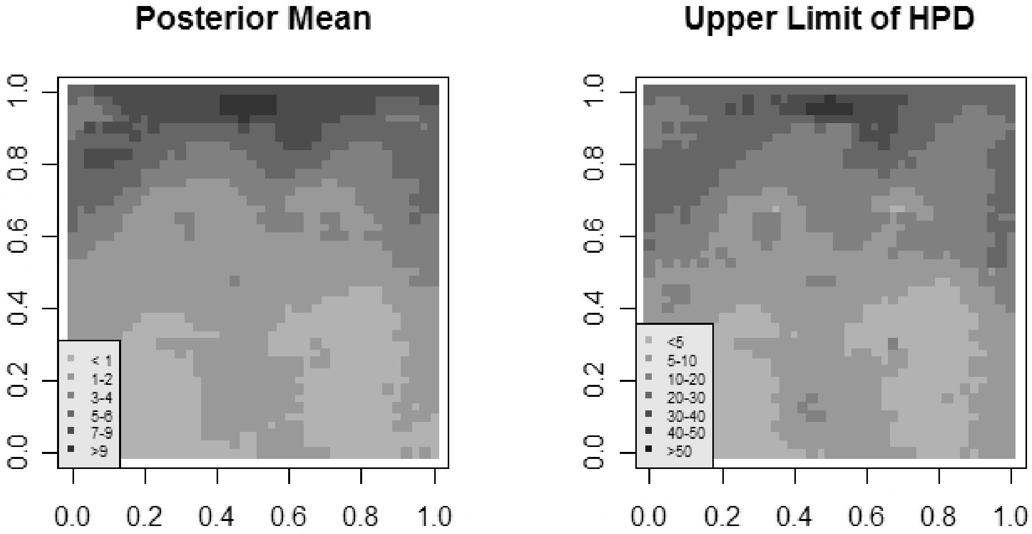


Fig. **6.11**: Posterior mean and upper limit of 95% HPD for predicted CPB larvae per meter row in the observed field under the TSNC model. The lower limit of the 95% HPD is 0 for all locations and is not mapped here.

6.4 Results from the TSIBNC model

In this section we apply the TSINBC model (as described in Section 5.5) to the CPB data set. We employed the same priors as the TSIB model.

Figure **6.12** plots posterior samples of the covariance and regression parameters, and the posterior distributions for TSIBNC are summarized in Table **6.4** . The posterior means for parameters in the abundance part of the model (θ_Z, β_0, β_1) are consistent with those from all three models previously applied.

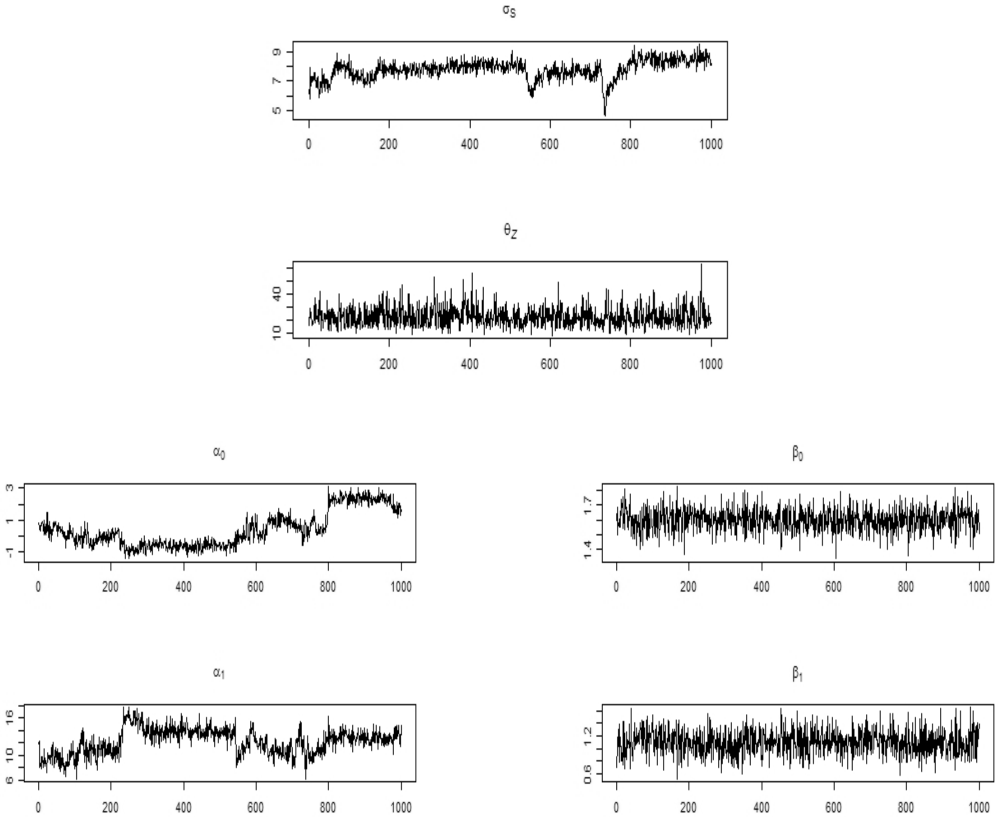


Fig. **6.12**: MCMC samples of all parameters for the CPB data set under the TSIBNC model.

Table **6.4**: Posterior estimates, Monte Carlo standard errors, and HPD for parameter estimates from the CPB data set under the TSIBNC model.

Parameter	Posterior Mean (Standard Deviation)	Monte Carlo Standard Error	95% HPD
σ_S	7.78 (0.64)	0.086	(6.58, 9.03)
θ_Z	22.36 (7.56)	0.258	(9.76, 37.66)
α_0	0.42 (1.10)	0.186	(-1.08, 2.59)
α_1	12.22 (2.10)	0.295	(7.84, 15.71)
β_0	1.60 (0.08)	0.002	(1.44, 1.74)
β_1	1.10 (0.20)	0.007	(0.70, 1.46)

Figure **6.13** maps the posterior mean of the *S* and *Z* samples, as well as mean expected incidence $\left(E\left(U\right)\right)$ and positive counts$\left(E\left(V\right)\right)$. With this covariance function, the *S* random effects are assumed to be independent and the cross-correlation between *S* and *Z* random effects was set to 0. In addition to the absence of spatial persistence in the map of *S* random effects, the removal of a cross-covariance between the two processes also seems to have increased the variance among the *S*. We note that the same trend was seen in applying this model to the simulated data. In Table **6.4** the posterior mean for σ_S is 7.78 compared to a posterior mean of 4.30 in the TSIB model (Table **6.2**). The posterior mean of the samples of *Z* random effects show spatial patterns similar to those in Figure **6.4** and Figure **6.7**. The maps of expected incidence $\left(E\left(U\right)\right)$ and abundance $\left(E\left(V\right)\right)$ are generally increasing functions of proximity to the north edge of the field, as observed earlier, with some localized patterns seen among the Z random effects.

Figure **6.14** maps the posterior mean and 95% HPD of predicted values of the semicontinuous response, similar to those in Figures **6.5** , **6.8** and **6.11** for the TSF, TSIB and TSNC models, respectively.

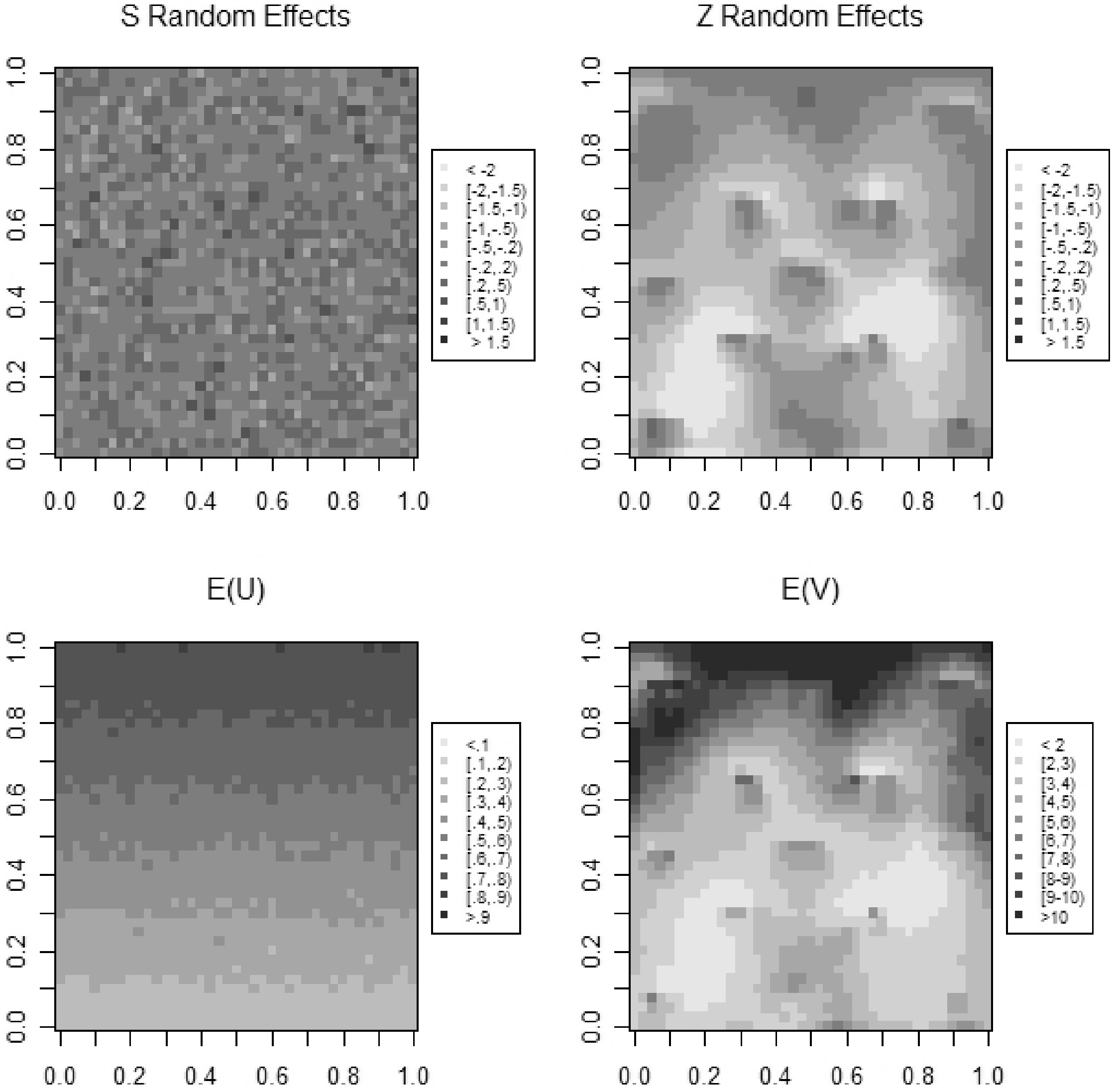


Fig. **6.13**: Predicted mean surface under the TSIBNC model.

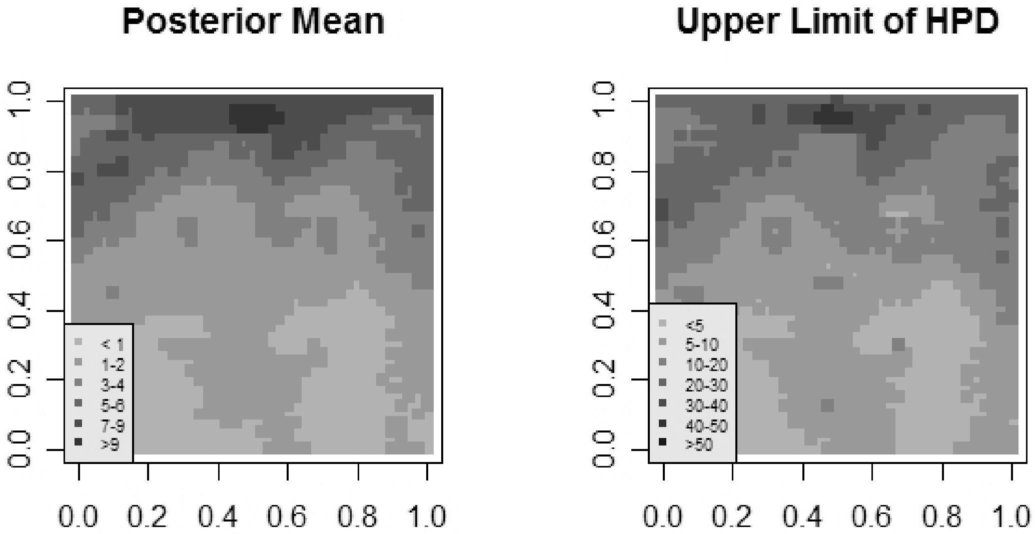


Fig. **6.14**: Posterior mean and upper limit of 95% HPD for predicted CPB larvae per meter row in the observed field under the TSIBNC model. The lower limit of the 95% HPD is 0 for all locations and is not mapped here.

6.5 Discussion

In this chapter we applied the two-part approach to modeling counts of CPB large larvae that have more zeros than can be accounted for by single discrete distribution. Two approaches that have been used to model data over space and time that contain excessive zeros are the two-part model (also called hurdle model) and the ZIP.

In this case the two-part approach is more appropriate than a ZIP model because it allows the modeling of the binary process separately from the count process. In biological settings incidence can be studied separately from abundance and simultaneous inference, such as prediction, is possible. It is appropriate when there is little chance of missing any items in the counts (Ver Hoef and Jansen 2007), so that a zero count can be taken as true absence of the species or phenomenon being counted.

We implemented the model in a geostatistical setting, assuming that the two processes arise from a bivariate stochastic process on a continuous spatial domain. We were able to generate smooth prediction surfaces by interpolating between unobserved locations. As in the application on simulated data (Chapter 5), we implemented the two-part model approach using four covariance structures. In all four covariance models, the posterior distributions for the parameters in the model for CPB abundance were comparable. The spatial dependencies in the prediction surfaces were mainly derived from the abundance part of the model, and therefore these were similar across models as well. The model for incidence detected only weak spatial dependence among the random effects and the random effects for the incidence and count parts do not appear to be correlated. However, our experience with the simulated data example has been that "absence of evidence is not evidence of absence". We have found that the random effects for the incidence part of the model are hard to quantify, and therefore it may be that in this case, there is undetected spatial persistence among these, and likewise that the process is correlated to the random effects from the abundance part of the model.

Table 6.5 summarizes the results from the four models when applied to the CPB data. The estimates of the parameters for the abundance part of the model (θ_Z, β_0, and β_1) were similar across the four covariance models, as it appears that the two-stage models captured the spatial dependencies as well as overall trend in this part of the model. However, as in the simulated data example, we observed that the HPD interval for the covariance parameter θ_Z was wider in the TSIB model (10.00, 48.09) compared to the other models. Again, this is probably due to an instability among the **Z** random effects when the model allows both correlation among the **Z** and cross-correlation to independent **S**.

Table **6.5**: Summary of HPD of MCMC samples for the different parameters under the four covariance models for the CPB data.

Parameter	95% HPD[a]			
	TSF Model	TSIB Model	TSNC Model	TSIBNC Model
σ_S	--	(2.61, 5.55)	--	(6.58, 9.03)
θ_S	(32.48, 114.03)	--	(43.78, 148.41)	--
θ_Z	(10.00, 33.56)	(10.00, 48.09)	(9.84, 39.27)	(9.76, 37.66)
ρ_{SZ}	(-0.20, 0.34)	(-0.27, 0.26)	--	--
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)

[a] The HPDs for the parameters in each column are generated under different model assumptions and are not directly comparable. In particular, the regression parameters for the logit part of the model (α_0, α_1) should be interpreted conditionally on their respective random effects, which vary between the different models.

With regards to the incidence part of the model, we found wider HPD intervals for α_0 when we reduce the spatial structure in the random effects for the incidence part. All the intervals include 0, but considerably wider under TSIB (-0.47, 0.91) and TSIBNC (-1.08, 2.59) compared to TSF (-0.31, 0.41) and TSNC (-0.24, 0.42). The posterior means and HPD intervals for α_1 are also greater in magnitude and width in TSIB and TSIBNC. Although we note that these estimates are generally not comparable because these were obtained conditional on random effects that varied depending on the covariance model, it appears that, as in the simulated data example, the estimates are more precise in the TSF and TSNC models.

The prediction maps for the semicontinuous variable Y generated under each model, shown in Figures **6.5** , **6.8**, **6.11** and **6.14** had similar spatial features. This is because these maps are mostly determined by the abundance part of the two-part model, and we have observed that the random effects and regression parameters for the abundance part of the model are similar across the four models. However, with respect to expected incidence $E(U)$, we note that predicted probability of incidence was smoother in the TSF and TSNC models (Figures **6.4** and **6.10**) compared to the TSIB and TSIBNC models (Figures **6.7** and **6.13**). Some spatial dependence, at shorter distances, was captured among Z random effects in the incidence part of the TSF and TSNC models via the assumption of an exponential covariance structure, and this resulted in smoother prediction maps.

Overall, we prefer the TSF and TSNC models over the TSIB and TSIBNC because both these models incorporate spatial dependence among random effects in both parts of the model. In this data set, we would recommend keeping the TSNC model because it retains the mechanism to capture spatial dependence in the incidence part of the model, and in this application this was enough to generate smoother maps and potentially more precise estimates of the regression parameters. The TSF model can be used even when the random effects vectors **S** and **Z** appear to be uncorrelated, because we can use the cross-correlation coefficient as a nuisance parameter to explore the effect of imposing cross-correlation on the posterior distributions. Although we found that spatial structure is more readily obtained from the abundance part, both parts of our two-stage model are equally important for interpretability, because the binary part is the sole predictor for incidence. For data with excessive zeros, it is particularly important that the binary model establish good estimates for predicting incidence, and this is where the TSF and TSNC performed better than the other two models.

Chapter 7

Summary and Future Work

We have considered the problem of modeling point-level spatial count data with a large number of zeros. We used a spatial generalized linear mixed model framework for the counts, employing a two-part approach to model incidence and abundance as separate but dependent processes, and utilized a bivariate Gaussian process model for characterizing the underlying spatial dependence. We fit this two-part model using a Bayesian approach via MCMC.

7.1 Summary

We implemented our approach on simulated data and a real data set from an ecological application. In addition to the two-stage full (TSF) model that includes dependence among random effects for counts, dependence among random effects for the binaries, and a cross-correlation between the two sets of random effects, we also tested the two-part model using simpler covariance functions for the random effects. The sub-models include independent random effects for the binary part (TSIB), removing the correlation between the two sets of random effects (TSNC), and a combination of the two (TSIBNC). We obtained prediction maps that are consistent with their known values (in the case of simulated data), and parameter estimates that are consistent with expected biological trends (in the case of the real data set). The two-part spatial modeling approach is computationally complex, particularly in sampling spatial random effects in the incidence part of the model. We found that spatial dependence was readily obtained from the model for positive counts (abundance) while the correlation parameters are only weakly identified by the binary outcomes in the incidence part of the model. The resulting predicted surface is still based on both predicted incidence and counts, but its spatial structure was mostly derived from the spatial dependence among random effects in the predicted counts. We also found that the regression parameters can be substantially affected by any collinearity between the covariates and the particular realization of the random effects in the data set, as we saw in the application to the simulated data.

Some of the spatial patterns among the binary (**S)** random effects can be captured using the TSF model because it has a mechanism for correlation among the **S** as well as cross-correlation to the random effects in the positive part (**Z**) which are also spatially dependent. The estimated regression coefficients for the positive counts are consistent across the three models. For the binary response, the TSNC and TSF model provided estimates that are closest to the true values and, in the case of the real data set, more precise estimates of the intercept. However, we note that in generalized linear mixed models, the regression parameters are interpreted conditional on the random effects. We observed that the spatial random effects for the binary outcome varied considerably among the four models, hence numerical differences in the regression parameters are difficult to intepret.

When the objective is both estimation and prediction, we recommend that the TSF and TSNC models be applied first. If we are studying a spatial phenomenon with excessive zeros that is compatible with a two-stage approach because the process determining incidence can be separated from the process determining abundance, then it is reasonable to assume that each set of spatial random effects are also spatially correlated and that there is cross-correlation between the two spatial processes. The results show that even if the spatial correlation from the binary part is only weakly identified, some spatial patterns can still be extracted via the correlation and cross-correlation structures. For prediction, it may be more parsimonious and computationally less intensive to assess spatial random effects solely from the abundance model and use TSIBNC, which assumes independent binary spatial random effects and no correlation between the binary and count random effects. On the other hand, if the focus is more on estimating the regression parameters, it may be advisable to apply the model while fixing the spatial dependence parameters to test whether the regression coefficients are consistent across a range of reasonable covariance parameters, much like the approach taken by Liang *et al.* (2008) in modeling dense point-level binary data.

We also note that the sampling plan should be adjusted according to the emphasis of the study. We designed a sampling plan with equal emphasis on estimating regression coefficients and covariance parameters, but if there is strong prior information regarding regression coefficients or if there is greater interest in spatial dependence, more samples should be allocated to closely-spaced locations.

We submit that the two-part approach, while often computationally complex, is useful when a spatially continuous stochastic phenomenon is observed to have excessive zeros, and where the process that generates the zero observations (incidence) is likely separate but potentially related to the process that determines abundance. The model can have different covariates for each part, allowing the covariates to impact each part of the response differently. Additionally, when it is reasonable to believe that these two mechanisms are related (not just because of shared covariates), there is a mechanism to relate the two processes using bivariate spatial random effects. We encountered computational difficulties in our MCMC algorithm when sampling random effects in the incidence part of the model, but we again point out that the cross-covariance function can be used essentially as a nuisance parameter to explore the other parameters of the model (e.g., regression coefficients). Finally, modeling the observations in an underlying spatially continuous stochastic process permits us to predict responses at arbitrary locations and generate a smooth predicted surface.

The two-part approach also facilitates comparison across different data sets because incidence is completely specified in the binary model and is separate from the distribution of positive counts. Because of this separation, the parameters have the same interpretation even for data sets that are not zero-inflated. Consistency in interpretation is also useful when the model is applied repeatedly to more than one data set, which may not all have excessive zeros. Examples are when modeling the same species over several fields or over time, or even for more than one species in the same field. For instance, the two-part model can still be applied even when a variable has a regular Poisson distribution. In a ZIP model where the zeros are assumed to arise from a mixture of distributions, it is not clear how the zeros will be allocated when the data set is no longer zero-inflated.

7.2 Future Work

While the two-stage model provides the flexibility to model data in accordance with a scientifically plausible data generating mechanism, it presents considerable computational challenges. The size of the covariance matrix increases rapidly with the size of the data set, and the joint specification of the model makes it susceptible to instability in the model fitting process. It would be interesting to explore the utility of alternative approaches for modeling large scale spatial data. Likelihood approximations such as those proposed by Caragea and Smith (2007) and Stein *et al*. (2004) could reduce computing intensity while still capturing the essential features of mean trends and spatial dependence. Gaussian Markov random field approximations to spatial Gaussian processes have also been proposed by Rue and Tjelmeland (2002). As an alternative to specifying the Gaussian process through its mean and covariance structure, Higdon (1998) and Higdon *et al*. (1999) use a process-convolution approach to develop non-stationary space-time models for datasets that are too large for straightforward kriging based methods.

In applying our models to simulated data we found that posterior means for the regression parameters, particularly for the abundance model, can be substantially affected by spurious collinearity between the known covariate and the particular realization of the random effects. While this is a chance feature of this data set, it is clear that in any particular data set, we should be aware that the random effects, as implicit fixed effects in every realization, can be confounded with the effect of the covariates. Reich *et al*. (2006) proposed diagnostics and methods to measure and alleviate the potential effects of collinearity between fixed effects covariates and random effects for both normal and non-normal observations with a CAR covariance structure for the random effects. An extension of this approach to a spatial generalized linear mixed models setting for point-referenced data would be of interest, as would a further extension to the two-stage modelling approach that we develop here.

Another concern related to the issue of increasingly complex models is the choice of prior distributions. Much of the work done to identify sensible priors was made using much simpler models, whereas these "sensible priors" are now used in more complicated models in an *ad hoc* fashion. There seems to be little guidance in the choice of priors, and often the use of uniform priors does not reflect lack of prior knowledge in these models.

We also did not explore the use of other covariance and cross-covariance functions for the spatial random effects. Various covariance and cross-covariance functions are used in earth science applications for multivariate spatial prediction. A variety of approaches for multivariable spatial data in a hierarchical Bayesian framework are also available, for instance, in Chapter 7 in Banerjee *et al*. (2004).

Finally, there seems to be no established criteria for determining whether a data set has excessive zeros, and at what point (say, what proportion of zeros) we can say that a more sensible model choice should account for excess zeros.

Bibliography

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, 341-355.

Anderson, T.W. (1958) *An Introduction to Multivariate Statistical Analysis*. Wiley and Sons, New York.

Armstrong, M. and Matheron, G. (1986) Disjunctive kriging revisited: Part I & II. *Mathematical Geology*, 18, 711-742.

Banerjee, S, Carlin, B.P. and Gelfand A.E. (2004) *Hierarchical modeling and analysis for spatial data*. Chapman and Hall, New York.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

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

Blom, P.E., Fleischer, S.J. and Smilowitz, J. (2002) Spatial and temporal dynamics of Colorado potato beetle (Coleoptera: Chrysomelidae) in fields with perimeter and spatially targeted insecticides. *Environmental Entomology,* 31, 149-159.

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

Caragea, P. C. and Smith, R. L. (2007). Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *Journal of Multivariate Analysis*, 98:1417–1440

Casella, G. and George, E.I. (1992) Explaining the Gibbs sampler. *The American Statistician*, 46, 167-174.

Chen, M.H., Shao, Q.M. and Joseph, G. (2000) *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

Chib, S. and Greenberg, E. (1995) Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49, 327-335.

Christensen, O.F., Möller, J., and Waagepetersen, R. (2000). Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte Carlo. Research Report R-00-2009, Department of Mathematical Sciences, Aalborg University.

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

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, 1–17.

Clayton, D.G. (1996) Generalized linear mixed models. In *Markov Chain Monte Carlo in Practice* (eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter), Chapman and Hall, New York.

Clayton, D.G. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671-681.

Cressie, N. (1989) The many faces of spatial prediction, in *Geostatistics*, Vol. 1, (ed. M. Armstrong). Kluwer, Dordrecht, 163-176.

Cressie, N. (1990) The origins of kriging. *Mathematical Geology*, 22, 239-252.

Cressie, N. (1993) *Statistics for Spatial Data*. Wiley and Sons, New York.

De Oliveira, V., Kedem, B., and Short, D.A. (1997) Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association*, 92, 1422-1433.

De Oliveira, V. (2000) Bayesian prediction of clipped Gaussian random fields. *Computational Statistics and Data Analysis,* 34, 299-314.

Diggle, P.J. (1996) Spatial analysis in biometry. In *Advances in Biometry* (eds. P. Armitage and H.A. David), Wiley and Sons, New York.

Diggle, P.J. (1997) Spatial and longitudinal data analysis: Two histories with a common future? *Lecture Notes in Statistics*, 122, 387-402.

Diggle, P.J., Harper, L. and Simon, S. (1997) A geostatistical analysis of residual contamination from nuclear weapons testing. In *Statistics for the Environment 3* (eds V. Barnett and K.F. Turkman), Wiley, Chichester, 89-107.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994) Longitudinal Data Analysis. Oxford University Press, New York:

Diggle, P.J., Ribeiro, P.J. and Christensen, O.F. (2002) An introduction to Model-Based Geostatistics. In *Spatial Statistics and Computational Methods* (J. Moller, Editor), Springer, New York, 43-86.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998) Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society Series C*, 47, 299-350.

Duan, N., Manning, W.G., Morris, C.N. and Newhouse, J.P. (1983). A comparison of alternative models for medical care. *Journal of Business and Economic Statistics*, 1, 115-126.

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

French II, N.M., Follett, P., Nault, B.A. and Kennedy, G.G. (1993) Colonization of potato fields in eastern North Carolina by Colorado potato beetle. *Entomologia Experimentalis et Applicata*, 68, 247-256.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *I.E.E.E. Transactions: Pattern Analysis and Machine Intelligence,* 12, 609-628.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J (eds.) (1996) *Markov Chain Monte Carlo in Practice,* Chapman and Hall, New York.

Goovaerts, P. (1998) Ordinary cokriging revisited. *Mathematical Geology*, 30, 21-42.

Greene, W.H. (1994) Accounting for excess zeros and sample selection in Poisson and Negative Binomial regression models, Technical Report EC-94-10, Stern School of Business, New York University.

Greene, W.H. (1997) FIML Estimation of Sample Selection Models for Count Data, Working Paper EC-97-02, Stern School of Business, New York University.

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

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.

Heckman, J. (1974) Shadow prices, market wages and labor supply. *Econometrica*, 42, 679-694.

Heckman, J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475-492.

Heckman, J. (1979) Sample selection bias as a specification error. *Econometrica*, 47, 153-161.

Heilbron, D.C. (1994) Zero-altered and other regression models for count data with added zeros. *Biometric Journal*, 36, 531-547.

Hedeker, D. and Gibbons, R.D. (2005) *Applied Longitudinal Data Analysis*. John Wiley, New Jersey.

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.

Higdon, D., Swall, J., and Kern, J. (1999) Non-stationary spatial modeling. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A., editors, *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*, pages 761–768. Clarendon Press [Oxford University Press].

Hough-Goldstein, J.A. and Whalen, J.M. (1996) Relationship between crop rotation distance from previous potatoes and colonization and population density of Colorado potato beetle. *Journal of Agricultural Entomology*, 13, 293-300.

Hur, K., Hedeker, D., Henderson, W., Khuri, S. and Daley, J. (2003) Modeling clustered count data with excess zeros in health care outcomes research. *Health Services and Outcomes Research Methodology*, 3, 5-20.

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

Isaaks, E.H. and Srivastava, R.M. (1990) Applied Geostatistics. Oxford University Press, New York.

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*, 101:1537–1547.

Kitanidis, P.K. (1983) Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, 19, 909-921.

Kitanidis, P.K. (1986) Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research*, 22, 499-507.

Kitanidis, P.K. (1997) *Introduction to Geostatistics: Applications in Hydrogeology*. Cambridge University Press, New York.

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

Lashomb, J.H. and Ng, Y.S. (1984) Colonization by Colorado potato beetles, Leptinotarsa decemlineata (Say) (Coleoptera: Chrysomelidae), in rotated and nonrotated potato fields. *Environmental Entomology* 13, 1352-1356.

Legendre, P. and Fortin, M.J. (1989) Spatial pattern and ecological analysis. *Vegetatio*, 80, 107-138.

Leung, S.F. and Yu, S. (1996) On the choice between sample selection and two-part models. *Journal of Econometrics,* 72, 197-229.

Liang, S., Banerjee, S., Bushhouse, S., Finley, A.O. and Carlin, B.P. (2008) Hierarchical multiresolution approaches for dense point-level breast cancer treatment data. *Computational Statistics and Data Analysis*, 52, 2650-2668.

Manning, W.G., Duan, N., and Rogers, W.H. (1987). Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics*, 35, 59-82.

Manning, W.G., Morris, C.N., Newhouse, J.P, Orr, L.L., Duan, N., Keeler, E.B., Leibowitz, A., Marquis, K.H., Marquis, M.S. and Phelps, C.E. (1981) A Two-part model of the demand for medical care: Preliminary results from the health insurance experiment. In *Health, Economics, and Health Economics* (eds. J. van der Graag and M. Perlman) Elsevier North-Holland, Amsterdam, 103-124.

Mardia, K.V. and Marshall, R.J. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71, 135-146.

Mardia, K.V. and Watkins, A.J. (1989) On multimodality of the likelihood in the spatial linear model. *Biometrika*, 76, 289-295.

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Low-Choy, S.J., Tyre, A.J. and Possingham, H.P. (2005) Zero tolerance ecology: improving ecological inference bymodelling the source of zero observations. *Ecology Letters*, 8, 1235-1246.

Matheron, G. (1962) Traite de Geostatistique Appliquee, Tome I. Memoires du Bureau de Recherches Geologiques et Minieres, No. 14. Editions Technip, Paris.

Matheron, G. (1976) A simple substitute for conditional expectation: The disjunctive kriging. In *Advanced Geostatistics in the Mining Industry* (eds. M. Guarascio, M. David, and C. Huijbregts) Reidel, Dordrecht, 221-236.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2nd ed, Chapman and Hall, London.

Metropolis, N. Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087-1092.

Møller, J. (Ed.) (2003) *Spatial Statistics and Computational Methods. Series: Lecture Notes in Statistics*, Vol 173, Springer, New York.

Oliver, D.S. (2003) Gaussian cosimulation: Modelling of the cross-covariance. *Mathematical Geology*, 35,681-698.

Omre, H. (1987) Bayesian kriging– merging observations and qualified guesses in kriging. *Mathematical Geology*, 19, 25-38.

Omre H. and Halvorsen, K. (1989) The Bayesian bridge between simple and universal kriging. *Mathematical Geology*, 21, 767-786.

Olsen, M. (1999) A Two-part Random Effects Model for Semicontinuous Longitudinal Data. Unpublished PhD Thesis, Pennsylvania State University, University Park, PA.

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

Potts, J.M. and Elith, J. (2006) Comparing species abundance models. *Ecological Modelling*, 199, 153-163.

Rathbun, S.L. and Fei, S. (2006) A spatial zero-inflated poisson regression model for oak regeneration. *Environmental and Ecological Statistics*, 13, 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, 1197-1206.

Ribeiro, P.J. and Diggle, P.J. (1999a) Bayesian inference in Gaussian model-based geostatistics. Technical Report ST-99-08, Department of Mathematics and Statistics, Lancaster University, Lancaster, UK.

Ribeiro, P.J. and Diggle, P.J. (1999b). geoS: S-PLUS functions for geostatistical analysis. Technical Report ST-99-09, Department of Mathematics and Statistics, Lancaster University, Lancaster, UK.

Ridout, M., Demetrio, C., and Hinde, J. (1998) Models for count data with many zeros. In: *International Biometric Conference*, Cape Town, pp. 1–13.

Rivoirard, J. (1994) *Introduction to Disjunctive Kriging and Non-Linear Geostatistics*, Clarendon Press, New York.

Robert, C.P. and Casella, G. (1999) *Monte Carlo Statistical Methods*, Springer, New York.

Roberts, G.O. and Rosenthal, J.S. (2001) Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16, 351-367.

Rossi, R.E., Mulla, D.J., Journel, A.G.and Franz E.H. (1992) Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs*, 62, 277-314.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29(1):31–49.

Schabenberger, O and Gotway, C.A. (2005) *Statistical Methods for Spatial Data Analysis*. CRC Press,New York.

Schotzko, D.J. and O'Keeffe, L.E. (1989) Geostatistical description of the spatial distribution of Lygus hesperus (Heteroptera: Miridae) in Lentils. J*ournal of Economic Entomology*, 82, 1277-1288.

Schotzko, D.J. and Smith, C.M. (1991) Effects of host plant on the between-plant spatial distribution of the Russian wheat aphid (Homoptera: Aphididae). *Journal of Economic Entomology*, 84, 1725-1734.

Shimizu, K. and Iwase, K. (1987) Unbiased estimation of the autocovariance function in a stationary generalized lognormal process. *Communications in Statistics, Theory and Methods*, 16, 2145-2154.

Stein, M.L. (1999) *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York.

Stein, M.L. (1998) In discussion of Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998) Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society Series C*, 47, 326-343.

Stein, M. L., Chi, Z., and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 66(2):275–296.

Stein, M.L. and Corsten (1991) Universal kriging and cokriging as a regression procedure. *Biometrics*, 47, 575-587.

Tu, W. (2002) Zero-inflated data. In: El-Shaarawi, A.H., Piegorsch, W.W. (Eds.), *Encyclopedia of Environmetrics*. John Wiley and Sons, Chichester, pp. 2387–2391.

Warnes, J.J. and Ripley, B.D. (1987) Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, 74, 640-642.

Williams, L., Schotzko, D.J. and McCaffrey, J.P. (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, 983-995.

Vecchia, A.V. (1988) Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society Series B*, 50, 297-312.

Vecchia A.V. (1992) A new method of prediction for spatial regression models with correlated errors. *Journal of the Royal Statistical Society Series B*, 54, 813-830.

Vella, F. (1998) Estimating models with sample selection bias: A survey. *The Journal of Human Resources*, 33, 127-169.

Ver Hoef, J.M. and Barry, R.P. (1998) Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, 69, 275-294.

Ver Hoef, J.M. and Cressie, N (1993) Multivariable spatial prediction. *Mathematical Geology*, 25, 219-240.

Ver Hoef, J.M. and Jansen, J.K. (2007) Space-time zero-inflated count models of Harbor seals, *Environmetrics*, 18, 697-712.

Webster, R. and Oliver, M. (2001) *Geostatistics for Environmental Scientists*, John Wiley and Sons, New York.

Zeger, S.L. and Karim, M.R. (1991) Generalized linear models with random effects : a Gibbs sampling approach. *Journal of the American Statistical Association,* 86, 79-86.

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

Zimmerman, D.L. (1989) Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. *Mathematical Geology*, 21, 655-672.

VITA

Virginia F. Recta

Education

Ph.D. in Statistics May 2009
The Pennsylvania State University, University Park, PA
Thesis title: A Model-based Analysis of Semicontinuous Spatial Data
Advisers: Dr. James L. Rosenberger, Dr. Murali Haran

M.S. in Statistics April 1987
B.S. in Statistics April 1983
University of the Philippines, Laguna, Philippines

Selected Professional Experience

Mathematical Statistician June 2004 to present
Center for Veterinary Medicine,
Food and Drug Administration, Rockville, MD

Biostatistician March 2002 to June 2004
Department of Biostatistics
Human Genome Sciences, Rockville, MD

Graduate Teaching Assistant January-May 2001
Department of Statistics, PSU, State College, PA

Senior Graduate Consultant Summer 1997, 1998, 1999
Statistical Consulting Center
Department of Statistics, PSU, State College, PA

Graduate Research Assistant Spring/Fall 1997, 1998, 1999, 2000
Department of Entomology
College of Agricultural Sciences, PSU, State College, PA

Publications

Petri, M., Stohl, W., Chatham, W., McCune, W.J., Chevrier, M. Ryel, J., Recta , V., Zhong, and J., Freimuth, W. (2008) Association of Plasma B Lymphocyte Stimulator Levels and Disease Activity in Systemic Lupus Erythematosus. Arthritis and Rheumatism, 58 (8), 2453–2459.

Recta, V.F., Wynn, K., Devon, R., and Derr, J. (2003) Retaining first-year women in science and engineering through research internships. World Transactions in Engineering and Technology Education 2(2): 179-184.