A Clipped Latent-Variable Model for Spatially Correlated Ordered Categorical Data

Megan Dailey Higgs^{*,a,1}, Jennifer A. Hoeting^b

^aMontana State University, Department of Mathematical Sciences, Bozeman, MT,
59717-2400, United States

^bColorado State University, Department of Statistics, Fort Collins, CO, 80523-1877, United States

Abstract

We propose a model for a point-referenced spatially correlated ordered categorical response and methodology for estimation of model parameters. Models and methods for spatially correlated continuous response data are widespread, but models for spatially correlated categorical data, and especially ordered multi-category data, are less developed. Bayesian models and methodology have been proposed for the analysis of independent and clustered ordered categorical data, and also for binary and count point-referenced spatial data. We combine and extend these methods to describe a Bayesian model for point-referenced (as opposed to lattice) spatially correlated ordered categorical data. We include extensive simulation results and show that our model offers superior predictive performance as compared to a non-spatial cumulative probit model and a more standard generalized linear model with spatial random effects. We demonstrate the usefulness of our model using a real-world example to predict ordered categories describing stream health within the state of Maryland.

Key words: Bayesian, ordinal, benthic IBI, generalized linear mixed models

1. Introduction

Ordered categorical data arise in a variety of scientific disciplines. The ordered categories may be of primary interest, or they may be recorded in an attempt to simplify data collection. We use the term “ordered categorical” to include both ordinal variables and interval variables, the latter having a meaningful numerical distance between two category labels. When the motivation for the categorization is simplification of a continuous variable, the categories are typically created using a set of (approximate) threshold values. While such a response may be easier and cheaper to obtain, the categorization complicates

*Corresponding author: Tel. 406-994-5350; Fax. 406-994-1789

URL: higgs@math.montana.edu (Megan Dailey Higgs)

¹M.D. Higgs

the statistical analysis since information regarding the continuous variable is clearly lost. However, an analysis of ordered categorical data can indirectly involve a latent (unobserved) underlying continuous variable. In fact, this concept creates a convenient platform on which to build models for ordered categorical data, both from an analytical and computational perspective. The goal of this paper is to describe an appropriate model for ordered categorical data collected at point-referenced locations over space. Such data are produced by a variety of research areas, such as ecology, epidemiology, and the social sciences.

Latent variable models for independent and correlated ordered categorical data has been an active area of research for over a decade. Bayesian inference is particularly attractive because the latent variables can naturally be incorporated as additional parameters using data augmentation. Methods for the analysis of independent data, most commonly using the ordinal probit model, are described by Albert and Chib (1993); Cowles (1996); Nandram and Chen (1996); Liu and Wu (1999); Chen and Dey (2000); Paquet et al. (2005); Roy and Hobert (2007), among others. The models for independent data have been extended to include correlation in the response(s), focusing on clustered and repeated measures data (Best et al., 1996; Chib and Greenberg, 1998; Chen and Dey, 2000; Ishwaran and Gatsonis, 2000; Li and Schafer, 2008).

Spatial models for point-referenced binary and count data have also received recent attention (e.g., Diggle et al., 1998; Gelfand et al., 2000; Christensen and Waagepetersen, 2002; Kammann and Wand, 2003; Christensen, 2004; Christensen et al., 2006; Eidsvik et al., 2006; Zhao et al., 2006). Most of these models remain within the convenient context of the one-parameter exponential family, and embed a Gaussian process within the framework of generalized linear mixed models (GLMMs). Such models have been coined generalized linear spatial models or generalized geostatistical models. Paciorek and Ryan (2005) provide a comparison of penalized likelihood and Bayesian models for point-referenced spatial logistic regression, focusing on computational issues. Generalized estimating equations (GEEs) and quasi-likelihood approaches for spatial generalized linear models are also available (Gotway and Stroup, 1997; Albert and McShane, 1995). Spatial and temporal models are more commonly described for areal (or lattice) binary, multinomial, and/or count data where inference is based on Markov random fields or autoregressive specifications (Vounatsou et al., 2000; Fahrmeir and Lang, 2001; Brewer et al., 2004; Zhu et al., 2005). Paquet et al. (2005) adopt a machine learning perspective for hierarchical ordinal regression. Bogaert (2002) introduces the use of Bayesian Maximum Entropy (BME) for spatial prediction of categorical variables, and also provides a thorough description of the severe limitations of indicator kriging (IK) and indicator co-kriging (ICK) when used to obtain the conditional probabilities needed to make categorical predictions. Namely, the estimated conditional probabilities obtained from IK do not sum to one since the probabilities of belonging to each category are obtained separately and those obtained from ICK are often negative.

An alternative approach to the generalized linear spatial model utilizes the concept of a clipped Gaussian random field, as described by De Oliveira (1997, 2000) for binary data. De Oliveira takes a step away from the GLMM approach

to more directly model the binary outcome, assuming a binary random field is created directly by clipping a Gaussian random field at a fixed threshold level, *i.e.* cut-point. This provides the theoretical basis for predicting the value of the binary random field based on a finite number of observations at point-referenced locations. Heagerty and Lele (1998) propose a similar strategy for binary data relying on a composite likelihood approach.

We propose a model we call the Clipped Gaussian model which extends De Oliveira’s work to multiple ordered categories. We assume a categorical random field is created by directly clipping an underlying latent Gaussian spatial process. A visual illustration of this concept is shown in Figure 1, with additional details given in Section 2. We offer an Bayesian algorithm for estimation of the model parameters. Our primary goal for the model is prediction. We assess predictions within the Bayesian decision theory paradigm, and evaluate predictions at new locations using hold-out data sets.

Figure 1: (Left) One realization from a continuous Gaussian random field. (Right) A four-category categorical random field created by clipping the continuous Gaussian random field with a vector of cut-points.

The remainder of this paper is organized as follows. In Section 2 we introduce the model, and in Section 3 we develop a Gibbs sampling algorithm (e.g., Givens and Hoeting, 2006) to facilitate Bayesian inference. Section 4 focuses on the details of prediction at new locations, Section 5 assesses predictive ability and estimation of parameters via simulation, and Section 6 provides an application of the model to ordinal data describing stream health through an “index of biotic integrity” (IBI) in Montgomery County, Maryland. Section 7 concludes with a brief discussion of the results and future work.

2. Model

Let $Y(\mathbf{s})$ be the categorical response at location $\mathbf{s}$ which can take on one of K categorical values. In what follows, we drop the dependence on $\mathbf{s}$ for notational convenience, so we define $Y = Y(\mathbf{s})$.

We assume a categorical random field, $\mathbf{Y}$, over the region of interest that is modeled by clipping, or thresholding, a continuous Gaussian spatial process, $\mathbf{W}$. The main goal is to map, or predict, an ordered multi-categorical value over a region of interest based on data collected at a finite number of locations within the region. The connection of the categorical random field to a Gaussian random field places the problem into the familiar context of spatial prediction for a continuous random field and geostatistical models, creating a workable mathematical framework (De Oliveira, 2000).

We define the latent Gaussian process, $\mathbf{W}$, as having a mean possibly depending on covariates and a valid spatial covariance matrix. That is,

$$\mathbf{W} \sim GP(\mathbf{X}'\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\gamma}))$$

where $\mathbf{X}$ is the covariate matrix, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\boldsymbol{\Sigma}$ is the covariance matrix. We use a no-nugget specification for $\boldsymbol{\Sigma}(\boldsymbol{\gamma})$ to write $\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \sigma^2 H(\boldsymbol{\psi})$ and $[H(\boldsymbol{\psi})]_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\psi})$, where ρ is a correlation function that produces a valid covariance matrix and depends on the spatial separation of the locations.

Using integer labels $1, \dots, K$ for the K categories, the categorical random field is assumed to be obtained from a realization of $\mathbf{W}$ as,

$$Y = \sum_{k=1}^K k I_{\{\theta_{k-1} < W \leq \theta_k\}},$$

where the θ 's are referred to as the cutpoints, and we assume that $\theta_0 = -\infty$, $\theta_1 = 0$, and $\theta_K = \infty$. The associated category probabilities arise from the deterministic relationship between $\mathbf{Y}$ and $\mathbf{W}$ such that

$$Pr(Y = k) = Pr(\theta_{k-1} < W \leq \theta_k). \quad (1)$$

The cumulative probabilities are similarly defined. Analogous non-spatial definitions are common (Agresti, 1999; Albert and Chib, 1993), and alternative parameterizations are described in Gelman and Hill (2007).

The unaugmented likelihood of the observed data, $\mathbf{y} = [y_1, \dots, y_n]^T$, is an n -dimensional integral of a multivariate Gaussian distribution with parameters from the latent Gaussian process:

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) &= \int_{\theta_{y_1-1}}^{\theta_{y_1}} \cdots \int_{\theta_{y_n-1}}^{\theta_{y_n}} \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} |H(\boldsymbol{\psi})|^{-1/2} \cdot \\ &\quad \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{w} - \mathbf{X}'\boldsymbol{\beta})' H(\boldsymbol{\psi})^{-1} (\mathbf{w} - \mathbf{X}'\boldsymbol{\beta}) \right\} d\mathbf{w}, \quad (2) \end{aligned}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\psi}, \sigma^2)$.

3. Inference

For inference, we appeal to data augmentation in order to incorporate the latent variable, $\mathbf{W}$, into the model. We estimate model parameters within the Bayesian paradigm using a Gibbs sampling algorithm paralleling that of De Oliveira (1997, 2000). However, unlike the binary case, we must sample the constrained cut-point parameters, thus increasing the level of complexity of the algorithm and leading to additional challenges in terms of convergence.

We factor the distribution of the unknown parameters and latent variable, given the observed data, as

$$f(\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma} | \mathbf{y}) \propto f(\mathbf{y} | \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(\mathbf{w} | \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}).$$

It is clear that the multidimensional integral in (2) is obtained by integrating $f(\mathbf{y} | \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ with respect to $\mathbf{w}$. However, use of the latent $\mathbf{w}$'s within the Gibbs sampler avoids the need for direct evaluation of this integral since the likelihood of $\mathbf{y}$ given the latent vector $\mathbf{w}$ is simply a product of indicator variables specifying whether w_i falls within the interval used to define the observed category, y_i , for $i = 1, \dots, n$:

$$f(\mathbf{y} | \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^n I_{\{\theta_{y_i-1} \leq w_i < \theta_{y_i}\}}.$$

This fact greatly simplifies the construction of the Gibbs sampling algorithm, which samples in turn from the complete conditional distributions of the unknown parameters and the latent variable.

3.1. Gibbs sampling algorithm

The algorithm iteratively proceeds through the following steps:

1. Update the spatial parameters, $\boldsymbol{\gamma}$, using the Metropolis-Hastings algorithm.
2. Update the latent spatial variable, $\mathbf{W}$, using the truncated multivariate Gaussian complete conditional distribution.
3. Update the cut-point parameters, $\boldsymbol{\theta}$, via reparameterization and a blocking algorithm.
4. Update the regression parameters, $\boldsymbol{\beta}$, using the multivariate Gaussian complete conditional distribution.

The details involved in these steps and the derivation of the complete conditionals are described in this section.

3.1.1. Spatial parameters, $\boldsymbol{\gamma} = (\sigma^2, \boldsymbol{\psi})$

The complete conditional distribution of the spatial parameters, $\boldsymbol{\gamma} = (\sigma^2, \boldsymbol{\psi})$, depends only on the latent variable, $\mathbf{W}$, and not on the observed data. It can be factored as

$$f(\boldsymbol{\gamma} | \mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}) \propto f(\mathbf{w} | \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(\boldsymbol{\gamma} | \boldsymbol{\beta}).$$

Until this point, it has been convenient to follow the standard parameterization for the spatial parameters of a geostatistical model, including both σ^2 and ψ . However, in the context of ordered categorical data, we are limited in the number of parameters we can estimate. For example, imagine a zero-mean normal distribution. By changing the cutpoints and the variance, identical categorical random variables can be obtained by infinitely many distributions. For example, consider $Pr(Y_i = 1)$ in the presence of a single covariate when $\theta_1 = 0$,

$$Pr(Y_i = 1) = Pr(W_i \leq 0) = \Phi\left(\frac{\beta_0 - \beta_1 x_i}{\sigma}\right). \quad (3)$$

As described by De Oliveira (2000) for binary data, it is clear from (3) that if σ and β are multiplied by the same constant, the probabilities do not change. This problem does not arise in non-spatial models, since the link function is typically specified as the cumulative distribution function (cdf) with fixed mean and variance, such as the standard logistic or standard normal.

We follow the convention of setting $\sigma^2 = 1$, though one could choose another value, as long as estimates of the regression parameters are then appropriately interpreted. That is, if data are generated under $\sigma^2 = 1$ and you fit a model assuming that $\sigma^2 = 2$, then the estimates of the regression parameters will be $\sqrt{2}$ times larger than the data generating values. We must remember that information is lost when going from a continuous variable to a categorical variable, and thus we cannot hope to estimate all the parameters specified in traditional spatial models for continuous data.

De Oliveira (1997, 2000) also describes near non-identifiability problems that can arise when estimating the smoothness (or roughness) parameter of the underlying correlation function using binary data. He points out that after clipping the continuous random field, the binary data contain no information about the parameter, even if we could observe the complete binary realization rather than at a finite number of locations. This is also a problem for multi-categorical data, and thus we restrict the set of correlation functions of interest to those using only one scale parameter to control the range of spatial correlation. For example, for this initial investigation we use the exponential correlation function so that $\gamma = (1, \phi)$ and $\rho(s_i - s_j; \phi) = \exp\{-d\phi\}$ where d is the distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$.

Our general strategy for prior specification is to find proper, yet relatively non-informative, distributions. That is, we seek to specify a prior distribution that spreads its density over a reasonable (practical) range of values while letting the data do most of the informing of the posterior distribution. We investigated several prior distributions for ϕ , and found the best convergence and most reasonable estimates by placing the prior on the theoretical effective range of the correlation function, rather than on ϕ itself. We denote the effective range as d_0 , and for the exponential correlation function it is approximately $3/\phi$. We then specify a uniform prior on the effective range from $d_{max}/100$ to d_{max} , where d_{max} is the maximum intersite distance in the data. This obviously assumes that the effective range is less than d_{max} . This prior places most of the probability density over reasonable values of ϕ and avoids the unreasonably large

values that are often obtained using other common priors. Another strategy that was adequate for some artificial data sets was that described by Banerjee et al. (2004), using a gamma distribution for ϕ with a mean specified using the effective range concept and d_{max} . For this specification, we found that setting the variance to “some large value” is much too general of statement, as small differences can lead to very different shapes of the distribution and that relatively large variances lead to obviously inappropriate prior distributions when visually examined, something also noted by Schmidt et al. (2007).

3.1.2. Spatial latent vector, $\mathbf{W}$

The assumption that $\mathbf{W}$ follows a Gaussian process with $\sigma^2 = 1$ and use of a one parameter correlation function implies that any vector of observations at n locations from one realization of the process has a multivariate Gaussian distribution,

$$\mathbf{W}_n \sim MVN_n(\mathbf{X}'\boldsymbol{\beta}, H(\phi)).$$

Let $\mathbf{w}_{-i}$ be the $\mathbf{w}$ vector with the i th element deleted. The full conditional distribution for each component of $\mathbf{W}$, for $i = 1, 2, \dots, n$, is given by

$$f(w_i|\mathbf{w}_{-i}, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi) \propto f(y_i|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi) f(w_i|\mathbf{w}_{-i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi).$$

Through the deterministic relationship between $\mathbf{Y}$ and $\mathbf{W}$, $f(y_i|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi)$ is simply an indicator function. The multivariate Gaussian distribution of $\mathbf{W}$ specifies a univariate Gaussian distribution for $f(w_i|\mathbf{w}_{-i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi)$ using the standard result found in texts such as Hocking (2003). It follows that $f(w_i|\mathbf{w}_{-i}, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi) \propto f(w_i|\mathbf{w}_{-i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi) I_{\{\theta_{y_i-1} \leq w_i < \theta_{y_i}\}}$, which is a truncated univariate Gaussian distribution which can be sampled using published techniques (Geweke, 1991; Rodriguez-Yam et al., 2004).

3.1.3. Cut-point parameters, $\boldsymbol{\theta}$

For the ordered multi-category case, we estimate $K - 2$ cut-point parameters since we parameterize the model by fixing the first cutpoint at zero and estimating β_0 as part of the $\boldsymbol{\beta}$ vector. We develop a method which utilizes the transformation proposed by Albert and Chib (1997) and the blocking algorithm proposed by Cowles (1996). The blocking method samples from the joint conditional distribution of the spatial latent variable, $\mathbf{W}$, and the cut-points, $\boldsymbol{\theta}$. The goal of the blocking is to reduce the correlation in the MCMC chain, thus accelerating convergence (Cowles, 1996). The strategy essentially translates into sampling the cut-points from their complete conditional distributions after marginalizing over the latent variable, $\mathbf{W}$. The below factorization of the joint conditional distribution provides insight into this step,

$$f(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y}, \boldsymbol{\beta}, \phi) = f(\mathbf{w}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi) f(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\beta}, \phi).$$

The de-constraint transformation of the cut-points expresses $\boldsymbol{\theta} = (\theta_2, \dots, \theta_{K-1})$ as an unconstrained and real-valued vector $\boldsymbol{\alpha} = (\alpha_2, \dots, \alpha_{K-1})$, where $\alpha_2 =$

$\log(\theta_2)$ and $\alpha_k = \log(\theta_k - \theta_{k-1})$ for $k = 3, \dots, K-1$. The inverse transformation is then given by $\theta_k = \sum_{i=2}^k \exp(\alpha_i)$. The target conditional distribution, assuming independence between $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and ϕ , is

$$f(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\beta}, \phi) \propto f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \phi)f(\boldsymbol{\alpha}).$$

The density $f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \phi)$ is the multi-dimensional integral given as the original likelihood function of interest in (2). The Metropolis-Hastings algorithm requires evaluation of both $f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$ and $f(\boldsymbol{\alpha})$ at current and proposed values of the transformed cut-point parameters. These values are readily available from evaluations of the multivariate Gaussian cdf and pdf, which can be accomplished with available software (Genz et al., 2009).

The transformation allows for use of a multivariate Gaussian prior distribution on $\boldsymbol{\alpha}$, $\boldsymbol{\alpha} \sim MVN_{K-2}(\mathbf{a}_0, \mathbf{A}_0)$. To specify the values of the elements in $\mathbf{a}_0$ and $\mathbf{A}_0$, we investigate the resulting prior on the original, untransformed cut-point parameters. For data with four categories, a diagonal specification of $\mathbf{A}_0$ with variance 0.5 or 1 results in a reasonable prior for $\boldsymbol{\theta}$. For larger diagonal elements, the mass is collected near the origin, giving a large amount of prior density to small values. Recall that the cut-point parameters are constrained to be greater than 0, since $\theta_1 = 0$. Smaller diagonal elements of $\mathbf{A}_0$ lead to accumulation of the mass farther from the origin. Specifying non-zero values for the off-diagonal elements of $\mathbf{A}_0$ resulted in very little difference in the posterior estimates, but slowed convergence of the algorithm for some values. Thus, after much investigation, we conveniently set $\mathbf{A}_0 = \mathbf{I}$ (further details can be found in Higgs, 2007).

Using Gaussian prior and proposal distributions for $\boldsymbol{\alpha}$ leads to the following transition probability defined by the Metropolis-Hastings ratio,

$$\min \left\{ \frac{f(\mathbf{y}|\boldsymbol{\alpha}^{cand}, \boldsymbol{\beta}, \phi)f(\boldsymbol{\alpha}^{cand})f_T(\boldsymbol{\alpha}^{cur}|\hat{\boldsymbol{\alpha}}, \psi^2 V, \nu)}{f(\mathbf{y}|\boldsymbol{\alpha}^{cur}, \boldsymbol{\beta}, \phi)f(\boldsymbol{\alpha}^{cur})f_T(\boldsymbol{\alpha}^{cand}|\hat{\boldsymbol{\alpha}}, \psi^2 V, \nu)}, 1 \right\}. \quad (4)$$

3.1.4. Regression parameters, $\boldsymbol{\beta}$

The full conditional distribution for the regression parameters, $\boldsymbol{\beta}$, is derived assuming prior independence between $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and ϕ

$$f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{w}, \boldsymbol{\theta}, \phi) \propto f(\mathbf{w}|\boldsymbol{\beta}, \phi)f(\boldsymbol{\beta}),$$

where $\mathbf{W}|\boldsymbol{\beta}, \phi$ has a multivariate Gaussian distribution. We also specify a conjugate multivariate Gaussian prior for $\boldsymbol{\beta}$, $\boldsymbol{\beta} \sim MVN(\mathbf{b}_0, \mathbf{B}_0)$. Collecting the terms involving $\boldsymbol{\beta}$, we obtain the standard linear models result,

$$\boldsymbol{\beta}|\boldsymbol{\theta}, \phi, \mathbf{y}, \mathbf{w} \sim MVN_p \left\{ (\mathbf{B}_0^{-1} + \mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1}, \tilde{\mathbf{B}}(\mathbf{B}_0^{-1} \mathbf{b}_0 + \mathbf{X}^T \Sigma^{-1} \mathbf{w}) \right\}.$$

We investigated the effect of many specifications on the convergence of the parameters for the case of a single covariate ($p = 2$) and finally selected the common specification of $\mathbf{b}_0 = \mathbf{0}$ and $\mathbf{B}_0 = \lambda^2 \mathbf{I}$, where λ^2 is set at a relatively large value to specify a vague prior. We noticed little difference in the results

or convergence behavior for values of $\lambda^2 > 20$ and thus typically used 20. This agrees with the findings of De Oliveira (1997). The results were very insensitive to the values of $\mathbf{b}_0$ when they were specified within the range of -10 to 10 , and thus we chose the convenient zero vector.

4. Prediction at new locations

A primary goal of our models, and of spatial models in general, is prediction at new locations. Thus, we evaluate the ability of our models to predict the categorical response at new, unobserved locations. We denote the vector of categorical response values at m new locations as $\mathbf{Y}_0 = (Y_{01}, Y_{02}, \dots, Y_{0m})$, and that of the latent variable as $\mathbf{W}_0$. Predictions at the new locations are made using the Bayesian posterior predictive distribution, $f(\mathbf{y}_0|\mathbf{y})$, and are denoted by $\tilde{\mathbf{Y}}_0$. We assess predictions within the Bayesian decision theory framework (De Oliveira, 1997, 2000; Berger, 1985), where prediction is based on the Bayesian expected loss, defined as the expected value of the loss function given the data. The loss function can be defined to suit the problem of interest, and the optimal Bayes predictor is found by minimizing the Bayesian expected loss.

4.1. Loss functions and Bayesian expected loss

We restrict our attention to a multi-category extension of the common additive binary loss function also employed by De Oliveira (1997, 2000). This additive loss function assigns zero loss to correct predictions and is given as,

$$L(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0) = \frac{1}{m} \sum_{j=1}^m \left[\sum_{k=1}^K \sum_{l \neq k} d_{kl} I_{\{Y_{0j}=k, \tilde{Y}_{0j}=l\}} \right], \quad (5)$$

where d_{kl} quantifies the loss incurred for predicting category l when the true category is k and $d_{kl} \geq 0$ for all k and l . The optimal Bayes predictor for (5) is simply the category with the largest probability. If we set all $d_{kl} = 1$, the loss function is the usual misprediction rate and the Bayesian expected loss (*BEL*) is

$$BEL = \frac{1}{m} \sum_{j=1}^m \sum_{k \neq \tilde{Y}_{0j}} Pr(Y_{0j} = k|\mathbf{y}). \quad (6)$$

4.2. The Bayesian posterior predictive distribution

To make predictions, we must estimate the conditional probabilities, $Pr(Y_{0j} = k|\mathbf{y})$, for $k = 1, 2, \dots, K$. These probability estimates are easily obtained from the MCMC algorithm, using the posterior predictive distribution $f(\mathbf{y}_0|\mathbf{y}) = \int f(\mathbf{y}_0|\boldsymbol{\eta})f(\boldsymbol{\eta}|\mathbf{y})d\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ represents the vector of all unknown parameters. Because we cannot directly sample from the posterior predictive distribution of $\mathbf{Y}_0$, we again augment the set of unknown parameters with $\mathbf{W}_0$, the unknown

latent values at the new locations. Thus, the desired augmented joint posterior distribution is $f(\mathbf{w}_0, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi | \mathbf{y})$, from which we can obtain draws from $f(\mathbf{w}_0 | \mathbf{y})$, the posterior predictive distribution needed to estimate the conditional probabilities.

As noted in De Oliveira (1997, 2000), the sampling can be divided into two stages due to the factorization,

$$f(\mathbf{w}_0, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi | \mathbf{y}) = f(\mathbf{w}_0 | \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi, \mathbf{y}) f(\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi | \mathbf{y}).$$

Thus, we first use the algorithm described in Section 3.1 to obtain T draws (after a burn-in period) from the original posterior distribution, $f(\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi | \mathbf{y})$. Then, we use $(\mathbf{W}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \phi^{(t)})$ to draw $\mathbf{W}_0^{(t)}$ from $f(\mathbf{w}_0 | \mathbf{w}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \phi^{(t)})$, for $t = 1, \dots, T$. Given the defined prior distributions, $f(\mathbf{w}_0 | \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \phi)$ is conveniently multivariate Gaussian.

4.3. Estimating the conditional probabilities

Recall that to make predictions we must estimate the conditional probability that a new location falls in category k given the observed data, $Pr(Y_{0j} = k | \mathbf{y})$, for each $k = 1, \dots, K$. The rule defined by minimizing the Bayesian expected loss is then applied to assign a categorical value, $\hat{Y}_{0j}$, to a particular location, $\mathbf{s}_j$. For the *BEL* in (6), this is simply assigning the category with the largest estimated probability. We obtain $Pr(Y_{0j} = k | \mathbf{y})$ by calculating $Pr(\theta_{k-1} < W_{0j} \leq \theta_k | \mathbf{y})$ using Monte Carlo integration over T iterations of the chain:

$$Pr(Y_{0j} = k | \mathbf{y}) \approx \frac{1}{T} \sum_{t=1}^T I_{\{\theta_{k-1}^{(t)} \leq w_{0j}^{(t)} < \theta_k^{(t)}\}},$$

where T is the total number of iterations considered after a burn-in period.

4.4. Bayesian expected loss as a measure of predictive uncertainty

To assess the adequacy of our predictive model, we desire measures of prediction uncertainty. Bayesian expected loss (*BEL*) provides a useful measure toward this end (De Oliveira, 1997, 2000). In this context, the term uncertainty refers to the *decision* to assign a certain value as the prediction, and not to the variability in the probability estimates used to make the predictions. De Oliveira (1997, 2000) refers to the Bayesian expected loss of the optimal Bayes predictor for binary data as a global measure of uncertainty about an estimated map based on the prediction locations. For the multi-category case and loss function in (5), we plug the K estimated probabilities for a location into (6) to obtain what we term the estimated BEL ($\widehat{BEL}$). For the j th location, $\widehat{BEL}_j = \sum_{k \neq \hat{Y}_{0j}} \widehat{Pr}(Y_{0j} = k | \mathbf{y})$, which is the sum of the $K - 1$ conditional probabilities after excluding the predicted category. Thus a prediction is considered to be more uncertain as the total probability is more spread out among the K categories. A global measure of predictive uncertainty is obtained by averaging $\widehat{BEL}_j$ estimates over all locations (De Oliveira, 1997, 2000).

It should be noted that the estimated BEL measure involves only the predicted value and is not appropriate for model validation. For example, the model could allocate the largest probability to category 1, when the true category is actually 4. This scenario results in a deceptively small $\widehat{BEL}$ and is obviously bad behavior for a model; not only is it making incorrect predictions, but it is doing so with false confidence. Validation measures of predictive ability relying on true values must also be used, such as use of hold-out data or cross validation.

4.5. Validation measures

We discuss assessment in the context of a hold-out data set, but analogous quantities can be obtained via cross-validation. The use of a hold-out data set is computationally much easier for a simulation study or large data set. Mis-prediction rate (MPR) is defined as the proportion of incorrect predictions, which again is (5) when all $d_{kl} = 1$.

In the context of ordered categories, it is also useful to employ a measure of *absolute mis-prediction rate* (AMPR) to account for the “size” of the mis-prediction. We define AMPR as

$$AMPR = \frac{1}{m} \sum_{j=1}^m |Y_{0j} - \tilde{Y}_{0j}|,$$

given successive integer labels for the K categories. We jointly utilize information from all three measures, $\widehat{BEL}$, MPR, and AMPR, to make conclusions regarding a model’s predictive ability. Another useful technique that we do not report here is a *BEL* validation measure where $\widehat{BEL}$ is calculated using the true values rather than the predicted values (Higgs, 2007). This is an easy check for the bad behavior mentioned in Section 4.4.

5. Simulation

To explore the usefulness of our model under realistic scenarios, we simulated data under a variety of scenarios. Previous work on ordered categorical data typically includes little simulation, most likely due to the computational expense. There clearly remains a need for more investigation via simulation of models for correlated ordered categorical data and even spatial models for non-Gaussian data in general. For the simulation results discussed here, we fit the model to over 150 different data sets, created under four different sets of parameter values using one covariate.

5.1. Sampling design

We are interested in comparing results from different realizations over fixed sampling locations. That is, we are interested in assessing the performance of the model over possible realizations the spatial process for one fixed sampling design. The locations were chosen to have a relatively large number of small

Figure 2: Locations of $N = 175$ sampling sites used for all data sets in the simulation study. Open circles indicate the $m = 50$ hold-out locations used to evaluate prediction and solid circles indicate the remaining $n = 125$ training locations.

inter-site distances, while at the same time providing adequate coverage over the study area (Zimmerman, 2006). In general, combining the two objectives suggests the use of a cluster-type design with more coverage introduced around the edges of the study region. In our experience, for larger sample sizes we obtain similar coverage from random designs, and thus the set of locations was chosen at random over a 10 by 10 unit region (Figure 2).

5.2. Spatial correlation

The success of our model’s estimates and predictions will depend, in part, on the spatial correlation in the categorical random field. We seek sets of parameter values that vary the degree of smoothness, or alternatively patchiness, in the categorical data. This is *not* simply an exercise in changing the parameters of the correlation function defining the Gaussian random field as it is for continuous data. Correlation in the categorical random field involves not only the spatial parameters, but also the regression parameters, β , the covariates, $\mathbf{X}$, and the cut-point parameters, θ . For illustration, Figure 3 compares the correlation in the categorical random field to that of the continuous random field for different values of a constant mean at a fixed distance and value for ϕ . For simplicity, the calculations use the integer labels for the categories, assuming they are meaningful as they would be for interval data. Notice that the correlation is always less than in the continuous random field. We use the results of such investigations of the second-order structure (details in Higgs, 2007) to set the values of the intercept and cutpoint parameters to obtain different levels of correlation in the categorical random field, along with different category compositions. We fixed $\beta_1 = 1$ for ease of comparison, and restrict our attention to the

Table 1: True parameter values defining the four scenarios used in the simulation study.

Scenario	β_0	β_1	θ_1	θ_2	θ_3	ϕ	σ_{true}^2
A	-0.5	1	0	1	2	0.5	1
B	2	1	0	1	2	0.5	1
C	-0.5	1	0	0.5	1.5	0.5	1
D	-0.5	1	0	0.5	1.5	0.5	3

exponential correlation function with $\phi = 0.5$. This results in an effective range of approximately 6, which is close to half the maximum inter-site distance.

Figure 3: Correlation between two locations at a distance of 1.41 using $(\theta_1, \theta_2, \theta_3) = (0, 1, 2)$ and $\phi = 0.5$. The horizontal lines indicate the correlation in the continuous random field for the specified value of ϕ .

5.3. Simulation scenarios

We create data under four parameter scenarios (Table 1) for a response with with four categories, labeled with integers 1 through 4 (Table 2). We use comparisons of the four scenarios to investigate our method's ability to produce accurate predictions and reasonable parameter estimates.

The model is fit using the algorithm detailed in this paper and implemented using R software (R Development Core Team, 2007). Prior distributions are specified as previously described and we use the same priors and starting values for all scenarios. The algorithm is generally insensitive to reasonable starting values with the specified priors. For each data realization, we ran the MCMC

Table 2: Observed percent of observations in each category for parameter scenarios defined in Table 1 averaged over realizations.

Scenario	1	2	3	4
A	58	20	14	8
B	12	19	21	48
C	62	12	15	11
D	64	8	14	14

Table 3: Simulation results for mis-prediction rate (MPR), absolute mis-prediction rate (AMPR), and estimated Bayesian expected loss (BEL) for the hold-out set of $m = 50$ new locations for Scenarios A, B, C, and D. The results in each row of the table are summarized over 30 realizations.

Scenario	MPR			AMPR			BEL		
	Min	Avg.	Max	Min	Avg.	Max	Min	Avg.	Max
A	0.14	0.37	0.52	0.16	0.47	0.70	0.22	0.31	0.46
B	0.24	0.47	0.62	0.36	0.62	0.86	0.28	0.38	0.47
C	0.24	0.35	0.58	0.40	0.61	1.08	0.17	0.30	0.42
D	0.26	0.43	0.58	0.38	0.62	1.06	0.25	0.37	0.48

chain for 3000 iterations, which is the same number used by De Oliveira (1997, 2000). Posterior summaries for prediction and estimation are based on the last 2500 iterations of the chain. Ideally, longer and multiple chains would be run, however we were limited by computational speed.

5.4. Prediction results

Table 3 gives the minimum, average, and maximum values over all realizations for mis-prediction rate (MPR), absolute mis-prediction rate (AMPR), and the estimated Bayesian expected loss (BEL) for $\sigma_{fix}^2 = 1$. The minimum and maximum values are included to flag the existence of realizations that result in particularly good or bad predictions.

Looking at three measures, the model appears most effective in terms of prediction on data generated under Scenario A, with low AMPR values indicating that the mispredictions were rarely more than one category away from the truth. This was an interesting and unexpected result given that the correlation among the categorical response values generated under A is lower than that of B and C for most values of the covariate. We initially suspected that the predictive success was simply a consequence of the large proportion of observations in the lowest category, however this conclusion is not consistent with the comparable results from Scenarios B and C.

The only difference between C and D is the value of σ_{true}^2 , such that $\sigma_{fix}^2 = \sigma_{true}^2$ for C and not for D. Scenario C data reveal slightly better MPRs, though AMPRs are nearly equivalent. However, when $\sigma_{fix}^2 = 2$, the predictions for Scenario D do improve so that MPRs are nearly identical and the AMPRs are actually slightly better for D. A similar reduction is not observed for $\sigma_{fix}^2 = \sigma_{true}^2 = 3$ for Scenario D. Thus, there is slight evidence that greater predictive success may accompany fixed values of σ^2 that are closer to the data-generating value.

Interestingly, Scenario C reveals the lowest values for all three measures of BEL, meaning that it resulted in higher maximum probabilities and thus greater certainty in the predictive decision. However, coupled with the relatively large AMPR values, it's clear that predictions for Scenario C data were often made with false certainty. The main difference between A and C data is the proportion of observations in second category.

An important observation from this investigation of artificial data is that the variability among realizations generated under the same scenario is nearly as great as the variability among scenarios. Sets of parameter values were chosen to produce visually distinguishable realizations while at the same time maintaining consistency in the values for the purpose of comparison. Despite this, when the data were ultimately randomly generated, it was often impossible to visually distinguish among realizations generated under different scenarios. This investigation demonstrates the applicability of the model to a great variety of potential data sets with different degrees of patchiness and category compositions. That being said, rather than averaging over many realizations, we now take a closer look at individual realizations in an attempt to identify attributes of data realizations for which the model is particularly successful, or unsuccessful.

Not surprisingly, less patchy realizations (more smooth) were generally accompanied by better predictions than those from more patchy (less smooth) realizations. The lowest minimum MPR and AMPR are observed on the same realization from Scenario A. The values are 0.14 and 0.16, respectively, meaning that 43 out of the 50 hold-out sites are correctly predicted, with only one prediction falling more than one category from the true value. The true category composition for the realization is (42, 5, 1, 2) and the predicted composition is (41, 7, 2, 0). This obviously represents a relatively smooth realization with a large proportion in the lowest category. Application of the model to artificial and real data indicates that the model does a particularly good job predicting values in the extreme categories. Another Scenario A realization resulting in relatively poor predictions had MPR and AMPR values of 0.44 and 0.56, respectively, corresponding to correct prediction of 28 out of 50 sites, with most mispredictions off by only one category. The true category composition for the realization is (22, 16, 8, 4) and the predicted composition is (18, 21, 8, 3).

De Oliveira (2000) reported mis-prediction rates of 0.16 and 0.19 for his relatively smooth simulated realizations, and 0.16 and 0.22 for the patchier realizations. These values are consistent with the values we obtain for our better predicted realizations. This is encouraging given that we are predicting a multi-

category response rather than binary, along with the fact that our realizations are typically more patchy. The range of our BEL measures are also consistent with De Oliveira (2000), ranging from 0.21 to 0.41. In fitting the model to over 150 different sets of data, on average 65% of the hold-out values were correctly predicted, with as many as 88% for some realizations. Importantly, the predictions were rarely more than one category away from the true value.

5.5. Parameter estimation results

We focus our discussion of estimation on β_1 since it is the parameter that will most likely be interpreted, and first discuss the effects of fixing σ^2 at a value different from the data generating one (truth). The effect on prediction of $\sigma_{fix}^2 \neq \sigma_{true}^2$ is barely noticeable, which is not surprising because the estimates of the regression parameters and cutpoints should simply be adjusted so that the model is estimating $\beta_{1,true} \times \sigma_{fix}$, rather than $\beta_{1,true}$. For example, if $\sigma_{true}^2 = 1$ and we fit a model with $\sigma_{fix}^2 = 2$, the estimates of the regression parameters and cutpoints must be multiplied by $\sqrt{2}$ to obtain correct category probability estimates. This results in posterior draws for β_1/σ_{fix} that are centered on $\beta_{1,true}/\sigma_{true}$. Figure 4 displays the posterior distributions of β_1/σ_{fix} over 30 different realizations, as well as the average and .025 and .975 quantiles of the posterior means. Notice that if $\beta_{1,true} = 1$, $\sigma_{true}^2 = 1$, and $\sigma_{fix}^2 = 2$, the posterior mean of β_1 is approximately $\sqrt{2}$, leading to a posterior mean of β_1/σ_{fix} of approximately 1.

Figure 4 also provides another glimpse of model behavior over many data realizations. The posterior distributions in Figure 4 clearly show that while the posterior distributions are centered near the true value on average, there are a few realizations for which the distribution almost excludes the data-generating value. However, using 95% posterior intervals to summarize the range of plausible values, it's clear that the model was successful at capturing the data-generating value for most realizations. Of the 60 model fits on the 30 realizations shown in Figure 4, 13 fail to include the true generating value, but none include zero, which could potentially be taken as a wrong conclusion of no relationship between the response and the covariate. Therefore, the model is clearly leading to the correct conclusion regarding the existence and direction of the relationship between the response and the covariate. This was true of all parameter scenarios.

6. Assessing stream health

The goal of the Clean Water Act of 1977 is “restoring and maintaining the chemical, physical, and biological integrity of the nation’s waters” (U.S. Environmental Protection Agency, 2003). Toward this end, streams and rivers must be monitored to both assess current state of health and provide reference values to compare with future observations. It is primarily the responsibility of individual states to conduct monitoring programs. Work initially focused on chemical integrity, but has more recently also emphasized biological integrity

Figure 4: Posterior distributions of β_1/σ_{fix} from 30 realizations generated under Scenario A fit using $\sigma_{fix}^2 = 1$ (left) and $\sigma_{fix}^2 = 2$ (right). The vertical dashed line represents the true value of the ratio ($\beta_{1,true}/\sigma_{true}$) and the arrow shows the average and .025 and .975 quantiles of the posterior means over the 30 realizations.

(U.S. Environmental Protection Agency, 2003). We demonstrate the utility of our model on biological stream data collected in Montgomery County, Maryland by Montgomery County’s Department of Environmental Protection and as part of the Maryland Biological Stream Survey (MBSS) for the Maryland Department of Natural Resources. A goal of the monitoring program is to rate stream areas according to their overall health as compared to reference streams.

One common measure of stream health is the *Index of Biotic Integrity* (IBI). Biotic integrity is defined by the U.S. EPA as “the condition of the biological communities (usually benthic macroinvertebrate and/or fish) of a waterbody based on comparison to a reference that is a relatively undisturbed system and represents the best quality to be expected for the ecoregion” (U.S. Environmental Protection Agency, 2007). They define IBI as a combination of metrics which describes community structure, function, and pollution sensitivity and view the index as providing a picture of overall ecological stream health. Separate IBIs are constructed for benthic macroinvertebrates (aquatic animals without backbones that dwell on the bottom of aquatic environments and are larger than 0.5 mm) and for fish and then categorized using a set of cut-point values based on reference streams to create the IBI *narrative* describing the overall health of a site as *poor*, *fair*, *good*, or *excellent* (U.S. Environmental Protection Agency, 2007).

6.1. Exploratory data analysis

The main goal of our analysis is to predict the categorical benthic IBI for new, unsampled locations within the study region. The streams of Montgomery County were densely sampled over this relatively small region (497 square miles or 1287 km²), providing an attractive data set for which to apply spatial models

Figure 5: Montgomery County is shown within the state of Maryland and the enlargement identifies the locations of the data collection stations.

(Figure 5). We utilized data at $N = 299$ stations over 8 years, from 1996 to 2003, with the majority of observations coming from 1996 to 1999. Montgomery County had a population of 855,000 in the year 2000, and lies adjacent to Washington D.C. We assume for the purpose of this analysis that the response is constant over this time span. The spatial distribution of the benthic IBI narratives for this sample is shown in Figure 6. For ease of visualizing the response over space, linear interpolation is used to provide a relatively smooth image over the study area, despite the fact that streams are obviously not found continuously over the plotted region.

Universal Transverse Mercator (UTM) coordinates are used for the analysis and are scaled to obtain distances in a numerically convenient range. This results in a maximum distance between stations (d_{max}) of 4.92, corresponding to a distance of 49.2 kilometers, or 30.6 miles. To informally explore the spatial correlation in the categorical data, a naive empirical semi-variogram for the benthic IBI scores was made using integer labels, $k = 1, \dots, 4$ (Figure 7). Notice that before accounting for any spatial trend, the spatial correlation appears to exhibit a large range relative to the maximum inter-site distance in the data. We explored the use of such variograms for many artificial data sets where we could compare the semi-variogram of the underlying continuous data to that of the categorical data. For a constant mean, the continuous and categorical empirical semi-variograms are surprisingly similar. On-going work is exploring the construction and use of variograms for exploratory data analysis for ordered categorical data.

Figure 6: Benthic IBI narrative for Montgomery County, Maryland. The integer value labels for the benthic IBI scores correspond to 1 = *Poor*, 2 = *Fair*, 3 = *Good*, and 4 = *Excellent*.

It is clear from Figure 6 that there is a spatial trend beyond what could be described by isotropic spatial correlation. We cannot, however, simply investigate the residuals after fitting a mean structure, as we would typically do with continuous data. After incorporating a covariate, the variance of observations at a distance of d apart depends on all parameters in the model, as well as the value of the covariate. We choose the covariate of *conductivity*, or *specific conductance*, as it appears to account for much of the trend. Conductivity was also chosen because it is easily obtained and generally useful for detecting contaminants in water since it is used as an indirect measure of the total dissolved solids (Murphy, 2007). Its value can indicate the presence of agricultural and road runoff and it can also be used as a surrogate measure for the geology and soil type at or upstream of the station Murphy (2007).

For model validation, we select a hold-out data set of $m = 100$, leaving a sample size of $n = 199$ from which to fit the model (Figure 8). A hold-out data set was used due to the relatively large sample size and the large computational burden associated with K -fold cross-validation.

6.2. Model specification

As in the simulation study, we use $\sigma_{fix}^2 = 1$ and specify priors as described in Section 3. We again adopt the one-parameter exponential correlation function, $\rho(d) = \exp\{-d\phi\}$. This isotropic correlation function depends only on the Euclidean distance between locations, which obviously does not take into account the networked structure and flow inherent of streams. A more appropriate application to stream data, which is beyond the scope of this paper, may be to

Figure 7: A naive empirical semi-variogram using integer labels for the four categories ($N = 299$).

Figure 8: The spatial layout of the stations used for this analysis. with triangles representing the hold-out data set of $m = 100$.

Table 4: Evaluation of predictive ability for the non-spatial, Generalized linear spatial model (GLSM), and the Clipped Gaussian (CG). Category compositions are shown along with mis-prediction rate (MPR), absolute mis-prediction rate (AMPR), and the estimated Bayesian expected loss (BEL).

Model	1	2	3	4	MPR	AMPR	BEL
Truth	22	24	42	12	-	-	-
Non-spatial	7	0	93	0	0.55	0.74	0.49
GLSM	11	26	63	0	0.43	0.46	0.30
CG	24	9	57	10	0.39	0.42	0.32

embed the methodology of Ver Hoef et al. (2006) and Cressie et al. (2006) into the framework of the clipped Gaussian random field.

Results are compared to a non-spatial cumulative probit model (abbreviated as “non-spatial” below) fit using maximum likelihood through the *polr* function in the MASS package of R statistical software (Venables and Ripley, 2002). The Bayesian version of the non-spatial cumulative probit model with our relatively non-informative priors gives nearly identical results, and can be fit with available software such as winBUGS (Lunn et al., 2000). We also compare the results to the generalized linear spatial model (GLSM) obtained by adding a random effect (with the same correlation structure as the CG model) to the usual cumulative probit model, and fit using winBUGS.

6.3. Prediction of benthic IBI

A primary goal of this analysis is to predict the categorical benthic IBI, as an indicator of stream health, for unobserved locations in Montgomery County. Toward this end, we evaluate predictions for the $m = 100$ hold-out locations. After tuning of the algorithm to obtain adequate convergence behavior, the results are based on one run of the MCMC algorithm for 8,000 iterations after a burn-in of 2000 iterations. Table 4 compares the predictive abilities of the three models, and clearly shows the superior prediction of the Clipped Gaussian model over the naive non-spatial cumulative probit model. The Clipped Gaussian model also outperforms the GLSM, particularly in the extreme categories. This may have important practical implications if, for example, only the sites predicted as *poor* are to be remediated.

6.4. Estimation of model parameters

We report and compare the estimates and posterior intervals for the most interesting parameters, β_1 and ϕ , among four models (Table 5). The GLSM₁ is the generalized linear spatial model parameterized such that the spatial process (second-stage) has zero mean and the first stage has mean $\mathbf{X}'\beta + \mathbf{W}$, while GLSM₂ is parameterized such that the spatial process has mean $(\mathbf{X}'\beta)$. We

Table 5: Posterior means of the coefficient for conductivity (β_1) and the scale parameter for the exponential correlation function (ϕ) with 95% posterior intervals for the Clipped Gaussian (CG) and generalized linear spatial models (GLSMs). An estimates and approximate 95% Wald’s confidence interval is given for the non-spatial cumulative probit model.

Model	β_1	ϕ
non-spatial	-0.63 (-0.79, -0.47)	-
GLSM ₁	-0.43 (-0.63, -0.22)	0.92 (0.48, 1.81)
GLSM ₂	-0.06 (-0.20, 0.05)	3.27 (1.35, 8.80)
CG	-0.10 (-0.18, -0.02)	1.68 (0.48, 2.43)

include the second parameterization because the CG model incorporates the mean structure directly into the spatial process.

Notice that all models produce a negative estimate for β_1 , indicating that the benthic IBI score tends to be lower for sites with higher measures of conductivity, as expected. The 95% Wald’s confidence interval for the non-spatial model and the 95% posterior interval for GLSM₁ exclude zero, giving convincing evidence for this negative relationship. The posterior means for the spatial models are closer to zero than the estimate from the non-spatial model. This may perhaps indicate that some of the spatial structure is explained through the correlation, rather than solely through the covariate. The estimate of β_1 for GLSM₂ is more similar to that of the CG than that of GLSM₁, showing that the level at which the mean structure is specified is important for inference.

We use the estimate of the correlation scale parameter, ϕ , to infer about the effective (or practical) range of spatial correlation. The effective range is commonly defined as the distance at which the correlation reaches 0.05 or where the variogram reaches 95% of its sill (Banerjee et al., 2004). The effective range for the exponential correlation function is approximately $3/\phi$. From the empirical semi-variogram shown in Figure 8, we visually estimate the effective range to be about 3.5, before accounting for the spatial trend explained by conductivity. The estimates of ϕ from CG and GLSM₁ both appear practically reasonable, giving an estimated effective range of 1.79 for CG and 2.5 for the GLSM₁ (accounting for a nugget fixed at 1, which is inherent in the GLSM model specification due to the probit link). The estimated effective range of 0.70 from GLSM₂ is closer to that of the CG model. The differences between CG, GLSM₁, and GLSM₂, have sparked further investigation into the behavior of the different parameterizations and the resulting implications for inference.

7. Discussion

In this work, we have described the Clipped Gaussian model for ordered categorical data, provided methodology for estimation, and demonstrated the success of the model for both simulated and real data sets. This model provides

a logical and important extension to the current literature surrounding spatial models for point-referenced ordered categorical data.

The application presented here provides a useful context for illustrating our method, however we do not comment on the appropriateness of benthic IBI as a measure of stream health (an issue of ongoing discussion in the ecological community). For new developments regarding the use of multimetric indices, such as IBI, for ecological health assessment, see Chiu et al. (2008).

An important area of future research is computational efficiency of the method, particularly for large data sets. Some techniques to consider include use of the spectral representation (Royle and Wikle, 2005; Paciorek and Ryan, 2005), marginal augmentation (Meng and van Dyk, 1999), approximate Bayesian methods (Gelfand et al., 2000; Christensen et al., 2006; Eidsvik et al., 2006), and non-parametric methods (Kammann and Wand, 2003).

On-going work involves understanding the theoretical and practical differences between the generalized linear spatial model (GLSM) (use of spatial random effects) strategy and our Clipped Gaussian strategy. The GLSM has a natural two-stage conditional specification with the first stage distribution defined by the probit link and the second stage defined by the assumed distribution for the spatial random effects. If the GLSM is reexpressed in its marginal form, the unit variance of the normal distribution associated with the probit link translates into a “nugget” term in the variance-covariance matrix that is fixed at one. This term is not inherent in the CG model, which is purposely specified with no nugget because it is not clear how its magnitude ultimately relates to properties of the categorical random field. The implications of this difference on the assumed spatial correlation structure of the ordered categorical variable are not well understood. For some discussion of the differences between a conditional (two-stage) and marginal (one-stage) specification in the context of continuous data see Banerjee et al. (2004), pages 131 and 148. Our results indicate that the Clipped Gaussian model has better predictive ability than the GLSM, especially when modeling a relatively smooth realization of a categorical random field, or when prediction of the extreme categories is of particular importance.

8. Acknowledgements

The work was supported by the STAR Research Assistance Agreement CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This work has not been formally reviewed by the EPA. The views expressed here are solely those of the author and STARMAP, the program they represent. EPA does not endorse any products or commercial services mentioned in this dissertation. Higgs research was also partially supported by the National Science Foundation IGERT under Grant DGE-0221595003 and Montana State University.

References

- Agresti, A., 1999. Modelling ordered categorical data: Recent advances and future challenges. *Statistics in Medicine* 18, 2191–2207.
- Albert, J., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88, 669–679.
- Albert, J., Chib, S., 1997. Bayesian methods for cumulative, sequential, and two-step ordinal data regression models. Tech. rep., Bowling Green State University.
- Albert, P. A., McShane, L. M., 1995. A generalized estimating equations approach for spatially correlated data: Applications to the analysis of neuroimaging data. *Biometrics* 51, 627–638.
- Banerjee, S., Carlin, B. P., Gelfand, A. E., 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC Press, Boca Raton, Florida, USA.
- Berger, J. ., 1985. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, New York, USA.
- Best, N. G., Spiegelhalter, D. J., Thomas, A., Brayne, C. E. G., 1996. Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society A* 159, 323–342.
- Bogaert, P., 2002. Spatial prediction of categorical variables: the Bayesian maximum entropy approach. *Stochastic Environmental Research and Risk Assessment* 16, 425–448.
- Brewer, M. J., Elston, D. A., Hodgson, M. E. A., Stolte, A. M., Nolan, A. J., Henderson, D. J., 2004. A spatial model with ordinal responses for grazing impact data. *Statistical Modelling* 4, 127–143.
- Chen, M.-H., Dey, D. K., 2000. A unified Bayesian approach for analyzing correlated ordinal response data. *Brazilian Journal of Probability and Statistics* 14, 87–111.
- Chib, S., Greenberg, E., 1998. Analysis of multivariate probit models. *Biometrika* 85, 347–361.
- Chiu, G. S., Guttorp, P., Liang, J., Khan, S. A., Westveld, A. H., 2008. A latent health factor index modelling approach via generalized linear mixed models, with application to ecological health assessment. Tech. rep., Department of Statistics and Actuarial Science, University of Waterloo.
- Christensen, O. F., 2004. Monte Carlo maximum likelihood in model-based geostatistics. *Journal of Computational and Graphical Statistics* 13, 702–718.

- Christensen, O. F., Roberts, G. O., Skold, M., 2006. Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics* 15, 1–17.
- Christensen, O. F., Waagepetersen, R., 2002. Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics* 58, 280–286.
- Cowles, M. K., 1996. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing* 6, 101–111.
- Cressie, N., Frey, J., Harch, B., Smith, M., 2006. Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics* 11, 127–150.
- De Oliveira, V., 1997. Prediction in some classes of non-Gaussian random fields. Ph.D. thesis, University of Maryland at College.
- De Oliveira, V., 2000. Bayesian prediction of clipped Gaussian random fields. *Computational Statistics and Data Analysis* 34, 299–314.
- Diggle, P. J., Tawn, J. A., Moyeed, R. A., 1998. Model-based geostatistics (with discussion). *Applied Statistics* 47, 299–350.
- Eidsvik, J., Martino, S., Rue, H., 2006. Approximate Bayesian inference in spatial generalized linear mixed models. Tech. rep., Department of Mathematical Sciences, Norwegian University of Science and Technology.
- Fahrmeir, L., Lang, S., 2001. Bayesian semiparametric regression analysis of multicategorical time-space data. *Annals of the Institute of Statistical Mathematics* 53 (1), 11–30.
- Gelfand, A. E., Ravishanker, N., Ecker, M. D., 2000. Generalized Linear Models: A Bayesian Perspective. Marcel Decker Inc, Ch. Modeling and inference for point-referenced binary spatial data, pp. 373–386.
- Gelman, A., Hill, J., 2007. *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge University Press, New York, New York, USA.
- Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T., 2009. mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-7. URL <http://CRAN.R-project.org/package=mvtnorm>
- Geweke, J., 1991. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. Tech. rep., University of Minnesota.
- Givens, G., Hoeting, J. A., 2006. *Computational Statistics*. John Wiley & Sons, Inc., Hoboken, New Jersey.

- Gotway, C. A., Stroup, W. W., 1997. A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics* 2, 157–178.
- Heagerty, P. J., Lele, S. R., 1998. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* 93, 1099–1111.
- Higgs, M. D., 2007. Clipped latent-variable spatial models for ordered categorical data. Ph.D. thesis, Colorado State University.
- Hocking, R. R., 2003. *Methods and applications of linear models: Regression and analysis of variance*. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Ishwaran, H., Gatsonis, C. A., 2000. A general class of hierarchical ordinal regression models with applications to correlated roc analysis. *The Canadian Journal of Statistics* 28, 731–750.
- Kamman, E. E., Wand, M. P., 2003. Geoadditive models. *Applied Statistics* 52, 1–18.
- Li, Y., Schafer, D. W., 2008. Likelihood analysis of the multivariate ordinal probit regression model for repeated ordinal responses. *Computational Statistics and Data Analysis* 52, 3474–3492.
- Liu, J. S., Wu, Y. N., 1999. Parameter expansion for data augmentation. *Journal of the American Statistical Association* 94, 1264–1274.
- Lunn, D. J., Thomas, A., Best, N., Spiegelhalter, D., 2000. Winbugs – a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 10, 325–337.
- Meng, X.-L., van Dyk, D. A., 1999. Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* 86, 301–320.
- Murphy, S. F., 2007. General information on specific conductance. Boulder Area Sustainability Information Network (BASIN).
URL <http://bcn.boulder.co.us/basin/data/BACT/info/SC.html>
- Nandram, B., Chen, M.-H., 1996. Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *Journal of Statistical Computation and Simulation* 54, 129–144.
- Paciorek, C. J., Ryan, L., 2005. Computational techniques for spatial logistic regression with large datasets. No. 32 in *Harvard University Biostatistics Working Paper Series*. The Berkeley Electronic Press.
- Paquet, U., Holden, S., Naish-Guzman, A., 2005. Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005. Springer Berlin / Heidelberg, Ch. Bayesian Hierarchical Ordinal Regression, pp. 267–272.

- R Development Core Team, 2007. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL <http://www.R-project.org>
- Rodriguez-Yam, G., Davis, R. A., Scharf, L. L., 2004. Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression. Tech. rep., Colorado State University.
- Roy, V., Hobert, J. P., 2007. Convergence rates and asymptotic standard errors for markov chain monte carlo algorithms for bayesian probit regression. *Journal of the Royal Statistical Society, Series B* 69 (4), 607–623.
- Royle, J. A., Wikle, C. K., 2005. Efficient statistical mapping of avian count data. *Environmental and Ecological Statistics* 12, 225–243.
- Schmidt, A. M., Conceicao, M. G., Moreira, G. A., 2007. Investigating the sensitivity of Gaussian processes to the choice of their correlation function and prior specifications. *Journal of statistical computation and simulation* to appear.
- U.S. Environmental Protection Agency, 2003. Introduction to the Clean Water Act. U.S. Environmental Protection Agency (EPA).
URL <http://www.epa.gov/watertrain>
- U.S. Environmental Protection Agency, 2007. Mid-Atlantic Integrated Assessment - The state of Maryland's freshwater streams. U.S. Environmental Protection Agency (EPA).
URL <http://www.epa.gov/maia/html/md-aq6.html>
- Venables, W. N., Ripley, B. D., 2002. *Modern Applied Statistics with S*. Springer, USA.
- Ver Hoef, J. M., Peterson, E., Theobald, D. M., 2006. Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics* 13, 449–464.
- Vounatsou, P., Smith, T., Gelfand, A. E., 2000. Spatial modelling of multinomial data with latent structure: an application to geographical mapping of human gene and haplotype frequencies. *Biostatistics* 1, 177–189.
- Zhao, Y., Staudenmayer, J., Coull, B. A., Wand, M. P., 2006. General design Bayesian generalized linear mixed models. *Statistical Science* 21 (1), 35–51.
- Zhu, J., Eickhoff, J. C., Yan, P., 2005. Generalized linear latent variable models for repeated measures of spatially correlated multivariate data. *Biometrics* 61, 674–683.
- Zimmerman, D. L., 2006. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics* 17 (6), 635–652.