A clipped latent variable model for spatially correlated ordered categorical data

Megan Dailey Higgs\(^a,\)\(^*\), Jennifer A. Hoeting\(^b\)

\(^a\) Montana State University, Department of Mathematical Sciences, Bozeman, MT, 59717-2400, United States
\(^b\) Colorado 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 inference. 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 simulation results and show that our model offers superior predictive performance as compared to a non-spatial cumulative probit model and a more standard Bayesian generalized linear spatial model. We demonstrate the usefulness of our model in a real-world example to predict ordered categories describing stream health within the state of Maryland.

**1. Introduction**

Ordered categorical data arise in a variety of scientific disciplines. The ordered categories may be the natural response variable of interest, or they may be recorded in an attempt to save cost at the time of 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. Often, the natural response variable is a continuous measurement, but is categorized at the time of data collection for logistical reasons. For example, substrate size in streams may be categorized into size classes, or percent cover in vegetation studies may instead be recorded as cover classes. We may never observe instances of the continuous random variable, but it is natural to envision the categorical data arising from a continuous distribution that is discretized (or clipped) via a set of predetermined threshold values. The categorization clearly results in a loss of information regarding any underlying continuous variable. However, analysis of ordered categorical data can be facilitated by appealing to the concept of such a latent (unobserved) continuous variable, even for problems where the existence of such a continuous variable is completely hypothetical (e.g., Agresti, 2002, page 277). The latent variable construct 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.

\(^*\) Corresponding author. Tel.: +1 406 994 5350; fax: +1 406 994 1789.
E-mail address: higgs@math.montana.edu (M.D. Higgs).

0167-9473/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.csda.2010.02.024
The use of latent variables in the construction of 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 in the model using data augmentation (Tanner and Wong, 1987). 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, focusing on clustered and repeated measures data (Best et al., 1996; Chib and Greenberg, 1998; Chen and Dey, 2000; Qiu et al., 2002; 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., 2009; Zhao et al., 2006). Most of these models remain within the convenient context of exponential family distributions, and embed a Gaussian process within the framework of generalized linear mixed models (GLMMs). The models proposed by Diggle et al. (1998) for binary and count data follow the spirit of Bayesian GLMMs, and have been coined generalized linear spatial models or generalized geostatistical models. Modeling cumulative probabilities for ordered categories is a logical extension to such methods. That is, a hierarchical model is defined by the inclusion of a spatial “random effect” term in the mean structure of a continuous latent variable, which is theoretically “clipped” to produce the categorical variable. We revisit and further discuss this modeling strategy in Section 5.

Other work also deserves mention. For example, 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). Machine learning and Bayesian Maximum Entropy have also been proposed for modeling spatial ordinal data (Paquet et al., 2005; Bogaart, 2002). Finally, the geosciences literature provides the truncated pluri-gaussian method as an approach to simulate from categorical random fields (e.g., LeLoc’h and Galli, 1997; Armstrong et al., 2003; Dowd et al., 2003; Emery, 2007), essentially clipping two or more potentially correlated Gaussian random fields to produce categorical random fields that are geologically realistic (LeLoc’h and Galli, 1997).

De Oliveira (1997, 2000) takes a step away from the GLMM approach to propose a unified Bayesian modeling strategy for binary data, 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 extend the work of De Oliveira to multiple ordered categories, proposing a model we term the clipped Gaussian model. We assume a categorical random field is created by directly clipping an underlying latent Gaussian random field. A visual illustration of this concept is shown in Fig. 1, with additional details given in Section 2. We offer a Bayesian MCMC algorithm for inference and a strategy for prediction of the ordered categorical response at new locations. We assess predictions within the Bayesian decision theory paradigm, and evaluate predictions at new locations using hold-out data sets.

The remainder of this paper is organized as follows: Section 2 formally introduces the model, Section 3 develops a Markov chain Monte Carlo (MCMC) algorithm to facilitate Bayesian inference, Section 4 focuses on the details of prediction at new locations, Section 5 provides an application of the model to ordinal data describing stream health through an “index of biotic integrity” (IBI) in Montgomery County, Maryland, and Section 6 concludes with a brief discussion of results and future work.
2. Model

Let \( D \) be a spatial region of interest, and \( Y(s) \) be an ordered categorical random field, defined continuously at every point \( s \in D \). That is, each point in the region of interest belongs to one of \( K \) ordered categories. For notational convenience, we assume throughout the paper that the labels for the \( K \) ordered categories are the successive integers from 1 to \( K \), and drop the dependence on \( s \) when referring to a finite number of locations. Further, we use \( Y \) to refer to the random variable at one location, \( s_i \), and \( Y \) as an \( n \times 1 \) vector of random variables from \( n \) locations. For observed values of random variables, we analogously use \( y_i \) and \( y \). Thus, the ordered categorical random variable at location \( i \) is denoted as \( Y(s_i) = Y_i \in \{ 1, 2, \ldots, K \} \), and the observed value of that random variable is denoted as \( y(s_i) = y_i \). For example, if \( y_i = 2 \), the \( i \)th observation is in the second category.

We envision the categorical random field, \( Y(s) \), is created by clipping, or thresholding, a continuous Gaussian random field, \( W(s) \), defined over the region of interest, \( D \). The main goal is to predict an ordered multi-categorical value over a region of interest based data collected at a finite number of point-referenced locations within the region. The connection of the categorical random field to the Gaussian random field places the problem into the familiar context of spatial prediction of interest based on data collected at finitely many locations. The Gaussian assumption for the latent distribution also places the model in the familiar framework of a cumulative probit model (De Oliveira, 2000).

The Gaussian assumption for the latent distribution also places the model in the familiar framework of a cumulative probit model (De Oliveira, 2000). We envision the categorical random field, \( Y(s) \), is created by clipping, or thresholding, a continuous Gaussian random field, \( W(s) \), defined over the region of interest, \( D \). The main goal is to predict an ordered multi-categorical value over a region of interest based data collected at a finite number of point-referenced locations within the region. The connection of the categorical random field to the Gaussian random field places the problem into the familiar context of spatial prediction of interest based on data collected at finitely many locations. The Gaussian assumption for the latent distribution also places the model in the familiar framework of a cumulative probit model (De Oliveira, 2000).

We define \( W(s) \) as a realization from a spatial Gaussian process with means depending on a linear combination of covariates. Then, for \( n \) (finite) locations,

\[
W \sim \text{MWN}_n(X\beta, \Sigma(y)),
\]

where \( X \) has dimension \( n \times p \), \( \beta \) is a \( p \times 1 \) including an intercept, and \( \Sigma(y) \) is the \( n \times n \) covariance matrix depending on parameter vector \( \gamma = (\sigma^2, \phi) \), separating the spatial variance parameter from those specified in the spatial correlation function. Thus, we write \( \Sigma(y) = \sigma^2 \gamma(\phi) \) and \( [\gamma(\phi)]_{ij} = \rho(s_i - s_j; \phi) \), where \( \rho \) is a correlation function producing a valid covariance matrix and depending on a measure of spatial separation between the locations \( s_i \) and \( s_j \). We use a no-nugget specification for \( \Sigma(y) \), since incorporation of a nugget parameter introduces non-identifiability and it is not clear at what value to fix such a parameter in this setting (Gelfand et al., 2000).

A realization of a categorical random field is obtained by clipping, or thresholding, \( W(s) \) using a vector of cut-points. Using the successive integers 1, 2, \ldots, \( K \) for category labels, we define the relationship between \( Y(s) \) and \( W(s) \) by

\[
Y(s) = \sum_{k=1}^{K} k I(y(s) \leq \theta_k),
\]

where the \( \theta \)'s are the ordered cut-points defined by

\[
-\infty = \theta_0 < \theta_1 < 0 < \theta_2 < \cdots < \theta_{K-1} < \theta_K = \infty.
\]

Note that \( \theta_1 \) is set to zero to avoid lack of identifiability with the intercept term in the mean of the latent process, though other parameterizations are also possible (e.g. Gelman and Hill, 2007, page 121).

The probability of the response at a single location falling into category \( k \) comes directly from the deterministic relationship between \( Y(s) \) and \( W(s) \), so that

\[
Pr(Y_i = k) = Pr(\theta_{k-1} < W_i \leq \theta_k).
\]

The cumulative probabilities are similarly defined, and analogous non-spatial cumulative probit definitions are common (e.g., Agresti, 1999; Albert and Chib, 1993).

2.1. Avoiding non-identifiability

Before discussing the resulting likelihood function, we make three simplifications for model identifiability. The first, as already mentioned, involves the common choice of setting \( \theta_1 = 0 \) to avoid non-identifiability with the intercept term. The second simplification is to set \( \sigma^2 = 1 \). Until this point, it has been convenient to follow the standard parameterization for the spatial parameters of a geostatistical model, including both \( \sigma^2 \) and \( \phi \). However, in the context of ordered categorical data, we are limited in the number of parameters we can estimate with relatively non-informative priors. The information contained in the categorical data pertains to the probability of being in a particular category and thus we cannot estimate both the variance of the (potentially hypothetical) latent continuous distribution and the cut-points (see De Oliveira (2000) for more discussion). Thus, we fix \( \sigma^2 = 1 \), following standard practice for non-spatial models, where the continuous latent variable is assumed to be standard normal (for a probit model) or standard logistic (for a logistic model).

The third simplification restricts the spatial correlation function to involve only one parameter, thus restricting \( \phi \) to a scalar. De Oliveira (1997, 2000) also describes near non-identifiability problems that can arise when estimating the smoothness (or roughness) parameter of the underlying spatial 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 a finite number of locations. This remains a problem for multi-categorical
data, and thus we restrict the set of correlation functions of interest to those defined with only one scale parameter to control the range of spatial correlation. For example, if the exponential correlation function is used, then \( \phi = \phi \) and \( \rho(s_i - s_j; \phi) = \exp(-d \phi) \) where \( d \) is the distance between locations \( s_i \) and \( s_j \).

2.2. Bayesian model specification

Recall that the observed data vector, \( y \), is a vector of \( n \) category labels taking on successive integer values from 1 to \( K \). We can thus use \( y_i \) to refer to the bracketing cut-points accompanying each observed value. For example, if \( y_i = 3 \), then the appropriate interval is \( \theta_{i-1} = \theta_2 \) and \( \theta_i = \theta_3 \). Using this notation, and incorporating the above simplifications, the augmented likelihood of the observed data, \( y = [y_1, \ldots, y_n]^T \), is an \( n \)-dimensional integral of a multivariate Gaussian distribution:

\[
F(y|\beta, \theta, \phi) = \int_{\theta_{1-1}}^{\theta_{1-n}} \cdots \int_{\theta_{n-1}}^{\theta_{n-n}} \left( \frac{1}{2\pi} \right)^{n/2} |H(\phi)|^{-1/2} \cdot \exp \left\{ -\frac{1}{2}(W - X\beta)'H(\phi)^{-1}(W - X\beta) \right\} dw.
\]  

To complete the Bayesian model specification, we define priors for all parameters: \( \phi, \beta, \) and the \( K - 2 \) unknown cut-points \( (\theta_2, \ldots, \theta_{K-1}) \). 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.

For \( \beta \), we assume a multivariate Gaussian distribution, \( \beta \sim \text{MVN}_p(\beta_0, B_0) \), where we set \( \beta_0 = 0 \) and \( B_0 = \lambda^2 I_p \), where \( I_p \) is the \( p \times p \) identity matrix, and \( \lambda^2 \) is a relatively large value to specify a vague prior. We investigated different prior specifications for the case of a single covariate \( (p = 2) \) and noticed little difference in the results or MCMC convergence behavior for values of \( \lambda^2 > 20 \) and thus typically used 20. This agrees with the findings of De Oliveira (1997). For a standardized covariate, the results are also insensitive to the values of \( \beta_0 \) between -10 and 10.

Use of an unconstrained re-parameterization of the cut-point parameters allows specification of a multivariate Gaussian prior. Details of the transformation and prior specification are given in Section 3.1. Prior specification for the spatial parameter, \( \phi \), should depend on the chosen correlation function. Here, we consider \( \phi \) as the parameter of the exponential correlation function as given in Section 2.1. We investigated several prior distributions with the goal of placing most of the prior density over values of \( \phi \) leading to reasonable values for the effective range, approximately \( 3/\phi \). For use of a gamma prior on \( \phi \), we find that care should be taken in specifying the hyperparameter values, as many default specifications suggested for continuous data can result in unreasonably large values (e.g., Banerjee et al., 2004, page 124). We also had success with placing the prior on the effective range, rather than directly on \( \phi \), assuming the effective range is uniform over \( d_{\text{max}}/100 \) to \( d_{\text{max}} \) (where \( d_{\text{max}} \) is the maximum inter-site distance in the data set). In this case, we still examine the resulting prior on \( \phi \). Specification of the prior for \( \phi \) should be problem dependent and we strongly recommend the good practice of visual examining the density before adopting the prior.

3. Inference

For inference, we appeal to data augmentation to incorporate the latent variable, \( W \). We make inference about model parameters within the Bayesian paradigm using a hybrid Gibbs–Metropolis Hastings sampling algorithm parallelizing 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

\[
f(W, \beta, \theta, \phi | y) \propto f(y | W, \beta, \theta, \phi) f(W | \beta, \theta, \phi) f(\beta, \theta, \phi).
\]

The incorporation of the latent variable simplifies the construction of the MCMC algorithm, which samples in turn from the complete conditional distributions of the unknown parameters and the latent variable. The algorithm iteratively proceeds through the following steps:
1. Update the spatial scale parameter, \( \phi \), using the Metropolis algorithm to sample from its complete conditional distribution.
2. Update the cut-point parameters, \( \theta \) and the latent spatial variable, \( W \), via a blocking algorithm and re-parameterization of the cut-points.
3. Update the regression parameters, \( \beta \), from their complete conditional distribution.

The details and complete conditional distributions involved in these steps are described in the remainder of this section.

3.1. Spatial parameter, \( \phi \)

The complete conditional distribution of the spatial parameter, \( \phi \), can be written to depend only on the latent variable, \( W \), and not also on the observed data. It can be factored as

\[
f(\phi | y, W, \beta, \theta) \propto f(W | \beta, \theta, \phi) f(\phi | \beta).
\]
To sample from the distribution, we use a Metropolis–Hastings (M–H) algorithm with a Uniform proposal distribution covering plausible values of \( \phi \) and centered at the current value of the parameter. This proposal distribution only requires evaluation of the complete conditional distribution, up to a constant, at the proposed and current values, since the symmetric proposal distribution drops out of the M–H algorithm (Givens and Hoeting, 2005). Thus, we must evaluate the function defined by multiplying the logarithm of multivariate normal probability density function evaluated at the current (or candidate) value by the logarithm of prior density of \( \phi \) evaluated at the current (or candidate) value.

3.2. Cut-point parameters, \( \theta \) and spatial latent vector, \( W \)

For the ordered multi-category case with \( \theta_1 = 0 \), we estimate \( K - 2 \) cut-point parameters. We combine the transformation proposed by Albert and Chib (1997) and the blocking algorithm for \( \theta \) and \( W \) proposed by Cowles (1996). The deconstrant transformation of the cut-points expresses \( \theta = (\theta_2, \ldots, \theta_{K-1}) \) as an unconstrained and real-valued vector \( \alpha = (\alpha_2, \ldots, \alpha_{K-1}) \), where \( \alpha_2 = \log(\theta_2) \) and \( \alpha_k = \log(\theta_k - \theta_{k-1}) \) for \( k = 3, \ldots, K - 1 \). The inverse transformation is then given by \( \theta_k = \sum_{i=2}^{k} \exp(\alpha_i) \).

The blocking method samples from the joint conditional distribution of the spatial latent variable, \( W \), and the transformed cut-points, \( \alpha \). The goal of the blocking is to reduce the correlation in the MCMC chain, thus accelerating convergence (Cowles, 1996). A multivariate algorithm could be used, but this is equivalent to using a univariate M–H algorithm to draw \( \alpha \) from \( f(\alpha|y, \beta, \phi) \), while separately drawing \( W \) from \( f(W|y, \beta, \alpha, \phi) \). The below factorization of the joint conditional distribution provides insight into this step,

\[
f(\alpha, W|y, \beta, \phi) = f(W|y, \beta, \alpha, \phi)f(\alpha|y, \beta, \phi).
\]

This strategy is particularly useful when the complete conditional distribution of one variable is easy to sample from and the marginal distribution of the other is easy to derive. This is essentially sampling the transformed cut-points from their complete conditional distribution after marginalizing over the latent variable \( W \) (Cowles, 1996). Thus, the target conditional distribution for updating the cut-points, assuming independence between \( \alpha \), \( \beta \), and \( \phi \), is

\[
f(\alpha|y, \beta, \phi) \propto f(y|\alpha, \beta, \phi)f(\alpha).
\]

The density \( f(y|\alpha, \beta, \phi) \) is the multi-dimensional integral given as the original likelihood function of interest in (1). The Metropolis–Hastings algorithm requires evaluation of both \( f(y|\alpha, \beta, \phi) \) and \( f(\alpha) \) 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 such as R (Genz et al., 2009). This does increase the computational burden by requiring evaluation of the multi-dimensional integral, however from preliminary investigations we found the improved convergence behavior was worth the extra computational burden. However, there are various changes that could be made to the MCMC algorithm at this point, including different transformations of the cut-points, that deserve further investigation and may be problem specific.

The transformation allows for use of a multivariate Gaussian prior distribution, \( \alpha \sim MVN_{K-2}(\alpha_0, \Sigma_0) \). To specify the values of the elements in \( \alpha_0 \) and \( \Sigma_0 \), we investigate the resulting prior on the scale of the original, untransformed cut-point parameters. Assuming a four-category problem with \( \theta_1 = 0 \), the prior in \( \alpha \) space can be written as,

\[
f(\alpha) = \frac{1}{\theta_2(\theta_3 - \theta_2)} f_\alpha(g^{-1}(\theta))
\]

where

\[
f_\alpha(g^{-1}(\theta)) = \frac{1}{2\pi |\alpha_0|^{-1/2}} \exp\left\{-\frac{1}{2} \left( g^{-1}(\theta) - \alpha_0 \right)^T |\alpha_0|^{-1} \left( g^{-1}(\theta) - \alpha_0 \right) \right\} I(0 < \theta_1 < \theta_2 < \theta_3)
\]

and \( g^{-1}(\theta) = (\log(\theta_2), \log(\theta_3 - \theta_2))' \).

We specify \( \alpha_0 \), the prior mean, by first choosing reasonable values for the cut-point parameters and then transforming those values to the \( \alpha \) space. Recall we are using unit variance for the continuous latent variable and setting \( \theta_1 = 0 \). Therefore, we must set the prior values for \( \theta_2, \ldots, \theta_{K-1} \) at ordered values greater than zero. Through graphical investigation, relatively non-informative hyperparameters were specified by finding quantiles (cut-points) that allocated approximately equal probabilities to the four categories. Some thought is needed regarding the mean value, but for a standardized covariate, we found little sensitivity for the four-category case comparing \((\theta_2, \theta_3) = (0.5, 1), (1, 2), \) and \((2, 4) \). De Oliveira (1997, 2000) suggests a strategy for use with binary data when some prior information regarding the relative size of the category probabilities can be elicited, which we also investigated for the multi–category case. It is relatively simple for the constant mean model, but quickly gets complicated when covariates are included.

For data with four categories, a diagonal specification of \( \Sigma_0 \) with variance 0.5 or 1 results in a reasonable prior for \( \theta \). For larger diagonal elements, the mass is collected near the origin, making small values more likely. Smaller diagonal elements of \( \Sigma_0 \) lead to accumulation of the mass farther from the origin. Specifying non-zero values for the off-diagonal elements of \( \Sigma_0 \) results in very little difference in the posterior distributions, but slows convergence of the algorithm for some values. Thus, we set \( \Sigma_0 = I \) (further details can be found in Higgs, 2007). In practice, we strongly suggest graphically exploring the prior on the original scale for different values of \( \alpha_0 \) and \( \Sigma_0 \).
The Metropolis–Hasting algorithm also requires a proposal distribution. Albert and Chib (1997) chose to match the proposal density, \( f_r \), to the target density, \( f(\alpha|\mathbf{y}, \beta, \phi) \). We also follow this idea, using a multivariate-\( t \) density with \( v \) degrees of freedom and parameters specified by locating the approximate mode, \( \hat{\alpha} \), and the inverse of the observed information matrix of \( \log(f(\alpha|\mathbf{y}, \beta, \phi)) \), denoted as \( \mathbf{V} \), using several Newton–Raphson iterations. Realizations are drawn from the multivariate-\( t \) by first sampling \( \lambda \) from a Gamma\(|\nu/2, \nu/2| \) distribution, \( \epsilon \) from a N\(|0, \psi^{-2}\mathbf{V}| \) distribution, and then calculating \( \alpha + \lambda^{-1/2} \epsilon \) (Albert and Chib, 1997). Thus, the algorithm requires two fixed tuning parameters, \( \nu \) and \( \psi^2 \), and is similar to a Langevin–Hastings algorithm (Givens and Hoeting, 2005).

To reduce computational time, we modify Albert and Chib’s algorithm to obtain \( \hat{\alpha} \) and \( \mathbf{V} \) for the first approximately 10 iterations and subsequently every 50th iteration. This simplification did not noticeably change the convergence behavior of the chain. Different values for the tuning parameters should be investigated for each application, but we found \( \psi^2 = 20 \) and \( \nu = 10 \) to result in acceptance of approximately half of the proposed values over a large variety of artificially generated data sets. The transition probability defined by the M–H ratio is then given as,

\[
\min \left\{ \frac{f(\alpha|\mathbf{y}, \beta, \phi)f(\alpha^{\text{rand}})f(\alpha^{\text{cur}}|\hat{\alpha}, \psi^2 \mathbf{V}, \nu)}{f(\alpha|\mathbf{y}, \beta, \phi)f(\alpha^{\text{rand}})f(\alpha^{\text{cur}}|\hat{\alpha}, \psi^2 \mathbf{V}, \nu)}, 1 \right\}.
\]

(2)

The current \( \alpha \) is then back-transformed to \( \theta \).

The latent spatial vector is updated under the assumption that \( \mathbf{W}(\mathbf{s}) \) is a realization from a Gaussian process with \( \sigma^2 = 1 \) and correlation defined by one parameter, \( \phi \). This implies that any vector of observations at \( n \) finite locations from one realization of the process has a multivariate Gaussian distribution,

\[
\mathbf{W} \sim \text{MVN}_n(X\beta, H(\phi)).
\]

Let \( \mathbf{W}_{-i} \) be the \( \mathbf{W} \) vector with the \( i \)-th element deleted. The full conditional distribution for each component of \( \mathbf{W} \) is given by

\[
f(W_i|\mathbf{W}_{-i}, \mathbf{y}, \beta, \theta, \phi) \propto f(y_i|\mathbf{W}, \beta, \theta, \phi)f(W_i|\mathbf{W}_{-i}, \beta, \theta, \phi),
\]

where \( f(W_i|\mathbf{W}_{-i}, \beta, \theta, \phi) \) is a univariate Gaussian distribution obtained using the standard result found in texts such as Hocking (2003). Through the deterministic relationship between \( \mathbf{Y} \) and \( \mathbf{W} \), \( f(\mathbf{y}|\mathbf{W}, \beta, \theta, \phi) \) is simply a product of indicator variables forcing each \( W_i \) to fall within the interval accompanying the observed category, \( y_i \):

\[
f(y_i|\mathbf{W}, \beta, \theta, \phi) = \prod_{i=1}^{n} I_{\{\theta_{y_{i-1}} \leq W_i < \theta_{y_i}\}}.
\]

Then, \( f(W_i|\mathbf{W}_{-i}, \beta, \theta, \phi) \propto f(W_i|\mathbf{W}_{-i}, \beta, \theta, \phi)I_{\{\theta_{y_{i-1}} \leq W_i < \theta_{y_i}\}} \) is a truncated univariate Gaussian distribution that can be sampled from with published techniques (Geweke, 1991; Rodriguez-Yam et al., 2004).

3.3. Regression parameters, \( \beta \)

The complete conditional distribution for the regression parameters, \( \beta \), is derived assuming prior independence between \( \beta, \theta, \) and \( \phi \).

\[
f(\beta|\mathbf{y}, \mathbf{W}, \theta, \phi) \propto f(\mathbf{W}|\beta, \phi)f(\beta),
\]

where \( \mathbf{W}|\beta, \phi \) is multivariate Gaussian. As described in Section 2.2, we specify a conjugate multivariate Gaussian prior, \( \beta \sim \text{MVN}(\beta_0, \Sigma_0) \). Collecting the terms involving \( \beta \), we obtain the standard linear models result,

\[
\beta|\theta, \phi, \mathbf{y}, \mathbf{W} \sim \text{MVN}_p\left(\mathbf{B}(\beta_0^{-1}\beta_0 + X^T\mathbf{H}(\phi)^{-1}\mathbf{W}), \mathbf{B}\right),
\]

where

\[
\mathbf{B} = (\beta_0^{-1} + X^T\mathbf{H}(\phi)^{-1}X)^{-1}.
\]

4. Prediction at new locations

An important goal of the Clipped Gaussian model, and of spatial models in general, is prediction at new locations. Thus, we evaluate the ability of our model to predict the categorical response at new, unobserved locations. We denote the vector of categorical variables at \( m \) new locations as \( \mathbf{Y}_0 = (Y_{01}, Y_{02}, \ldots, Y_{0m}) \), 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 \( \mathbf{Y}_0 \). 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. In this section, we first define the prediction problem in the context of common loss functions. We then describe the method for obtaining probabilities for each category at each location, ultimately needed to make predictions. Finally, we describe the use of an estimated Bayesian expected loss as a measure of predictive uncertainty.
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):

\[
L(Y_j, \tilde{Y}_j) = \frac{1}{m} \sum_{j=1}^{m} \left[ \sum_{k=1}^{K} \sum_{k \neq j} d_{kl}(Y_j = k, \tilde{Y}_j = l) \right].
\]

This loss function assigns zero loss to correct predictions, while \(d_{kl}\) quantifies the loss incurred for predicting category \(l\) when the true category is \(k\), with \(d_{kl} \geq 0\) for all \(k\) and \(l\). For all \(d_{kl} = 1\), the loss function is the usual misprediction rate (MPR), defined as the proportion of predictions that are incorrect, and the Bayesian expected loss (BEL) is

\[
\text{BEL} = \frac{1}{m} \sum_{j=1}^{m} \sum_{k \neq Y_j} \Pr(Y_j = k | y).
\]

For this situation, the optimal Bayes predictor is simply the category associated with the largest probability.

In the context of ordered categorical data, it is natural to also consider incurring greater loss for predictions that are farther from the truth. For example, for successive integer category labels, we could assign loss coefficients \(d_{kl} = |k - l|\). The resulting loss function defines the absolute misprediction rate (AMPR),

\[
\text{AMPR} = \frac{1}{m} \sum_{j=1}^{m} |Y_j - \tilde{Y}_j|,
\]

where \(Y_j\) is the true value at location \(j\) and \(\tilde{Y}_j\) is the prediction. We use AMPR, along with MPR, to validate the predictive ability of the model on hold-out data sets. We do not specify the optimal Bayes predictor from the second loss function (AMPR), but that is a logical and easy extension, predicting \(Y_j = k\) if \(\min(\text{BEL}_1, \text{BEL}_2, \ldots, \text{BEL}_K) = \text{BEL}_k\).

4.2. Obtaining the conditional probabilities

To make predictions using either of the above mentioned loss functions, we must first obtain estimates of the conditional probabilities, \(\Pr(Y_j = k | y)\), for \(k = 1, 2, \ldots, K\). These probability estimates are easily obtained from the MCMC algorithm, using the posterior predictive distribution \(f(Y_j | y) = \int f(Y_j | \eta) f(\eta | y) d\eta\), where \(\eta\) represents the vector of all unknown parameters. Because we cannot directly sample from the posterior predictive distribution of \(Y_j\), we again augment the set of unknown parameters with \(W_0\), the unknown latent values at the new locations, giving the desired joint posterior distribution, \(f(W_0, W, \beta, \theta, \phi | y)\). From this, draws are obtained from the posterior predictive distribution, \(f(W_0 | y)\), and used to estimate the conditional probabilities.

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

\[
f(W_0, W, \beta, \theta, \phi | y) = f(W_0 | W, \beta, \theta, \phi, y) f(W, \beta, \theta, \phi | y).
\]

Thus, we first use the algorithm described in Section 3 to obtain \(T\) draws (after a burn-in period) from the original posterior distribution \(f(W, \beta, \theta, \phi | y)\). Then we use \(f(W_0 | W, \beta, \theta, \phi)\) to draw \(W_0^t\) from \(f(W_0 | W^t, \beta, \theta, \phi)\) for \(t = 1, \ldots, T\), which is convenientely multivariate Gaussian.

Then, we obtain \(\Pr(Y_j = k | y)\) by calculating \(\Pr(\theta_{k-1} < W_{0j} \leq \theta_k | y)\) using Monte Carlo integration over \(T\) iterations of the chain:

\[
\Pr(Y_j = k | y) \approx \frac{1}{T} \sum_{t=1}^{T} I_{\theta_{k-1} < W_{0j}^t \leq \theta_k}.
\]

The prediction under the MPR loss function is simply made by choosing the category with the largest associated conditional probability.

4.3. Bayesian expected loss as a measure of predictive uncertainty

Along with the validation measures of MPR and AMPR requiring a hold-out data set or cross-validation, we explore the use of the Bayesian expected loss (BEL) to obtain a measure of prediction uncertainty (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. For the multi-category case and misprediction rate (MPR) loss function, we plug the \(K\) estimated probabilities for a location into (3) to obtain an estimated BEL (BEL). For the \(j\)th location, \(\hat{\text{BEL}}_j = \sum_{k \neq Y_j} \Pr(Y_j = k | y)\), the sum of the \(K - 1\) estimated probabilities after excluding that of 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 \(\hat{\text{BEL}}_j\) over all locations (De Oliveira, 1997, 2000).
Fig. 2. Montgomery County is shown within the state of Maryland and the enlargement identifies the locations of the data collection stations.

It should be noted that the \( \hat{BEL} \) measure involves only the prediction, not the true value, and thus is not appropriate for validation. For example, the model could allocate the largest probability to category 1, when the true category is 4, resulting in a deceptively small \( \hat{BEL} \); an incorrect prediction, accompanied by false confidence. This is obviously bad behavior for a model and thus it is also important to assess predictive ability by direct comparison of predictions to known true values. Such measures include MPR and AMPR, as previously defined. We jointly utilize information from all three measures to make conclusions regarding a model’s predictive ability. Another potentially useful technique is construction of a validation BEL, where BEL is calculated using the estimated probabilities and the true values rather than the predicted values (Higgs, 2007). Discrepancies between BEL and validation BEL are an easy check for the bad behavior mentioned above. We use a hold-out data set due to the computational burdens of cross-validation in this already computationally intensive setting.

5. 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” (US 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 (US 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.

A common measure of stream health is the Index of Biotic Integrity (IBI). Biotic integrity is defined by the US EPA as “the condition of the biological communities (usually benthic macroinvertebrate and/or fish) of a water body based on comparison to a reference that is a relatively undisturbed system and represents the best quality to be expected for the ecoregion” (US Environmental Protection Agency, 2007). IBI is defined as the sum of a combination of metrics describing community structure, function, and pollution sensitivity. It is viewed as a measure of overall ecological stream health, but the values themselves have no physical interpretation. Thus, to facilitate management, the IBI values are translated to an ordinal scale with levels poor, fair, good, or excellent. This four-category simplification of IBI is termed the IBI narrative and can be considered a site-level report card, a popular approach with governing bodies thus motivating interest to directly model the IBI narrative.

5.1. 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 km\(^2\)), providing an attractive data set for which to apply spatial models (Fig. 2). We utilized data at \( N = 299 \) stations collected over 8 years (1996 to 2003), with the majority of observations between 1996 and 1999. Montgomery County had a population of 855,000 in the year 2000, and Washington DC lies immediately southeast of the county. We assume for the purpose of this analysis that the response is constant over this time span. The observed spatial distribution of the benthic IBI narratives for this sample is shown in Fig. 3. 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 ($d_{\text{max}}$) of 4.92, corresponding to a distance of 49.2 km, 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, $k = 1, \ldots, 4$ (Fig. 4). 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 from the underlying continuous data to that from the categorical data. For a constant mean, the continuous and categorical empirical semi-variograms are strikingly similar, particularly for visual estimation of an effective range.

It is clear from Fig. 3 there is a spatial trend beyond what could be described by isotropic spatial correlation. It is not appropriate, however, to simply investigate the residuals after fitting a mean structure, as we would 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 its value can indicate the presence of agricultural and road runoff or it can be used as a surrogate measure for the geology and soil type at, or upstream of, the station (Murphy, 2007).

For model validation, we randomly select a hold-out data set of $m = 100$, leaving a sample size of $n = 199$ from which to fit the model. 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.

5.2. Model specification and comparisons

We specify priors as described in Section 2.2, fix $\sigma^2 = 1$, and 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
Comparison of models used for stream health assessment. In the Linear Model column, $\epsilon$ is an $n$-vector of independent errors and the other components are defined similarly to the components of the CG model. $I_n$ is the $n \times n$ identity matrix. All models are based on a cumulative probit framework linking the categorical response, $Y$, to a normally distributed latent random variable $Z$ ($W$ for the CG model). Note that the CG model directly links the spatially correlated random variables, $W$, to the discrete response $Y$ via the deterministic function. The other models discretize (conditionally) independent random variables, $Z$. The GLSM models only differ in the level at which $\mathbf{X}_B$ is specified, resulting in different conditional distributions in the MCMC algorithm. See text for additional model details.

<table>
<thead>
<tr>
<th>Model</th>
<th>Deterministic function</th>
<th>Linear model</th>
<th>Distributions for latent variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-spatial</td>
<td>$Y_i = \begin{cases} k &amp; \text{if } \beta_{k-1} &lt; Z_i \leq \beta_k \ 0 &amp; \text{otherwise} \end{cases}$</td>
<td>$Z_i = \mathbf{X}_B + \epsilon$</td>
<td>$Z_i \sim \text{MVN}(\mathbf{X}_B, I_n)$</td>
</tr>
<tr>
<td>GLSM$_1$</td>
<td>Same as above.</td>
<td>$Z_i = \mathbf{X}_B + W_i + \epsilon$</td>
<td>$Z_i</td>
</tr>
<tr>
<td>GLSM$_2$</td>
<td>Same as above.</td>
<td>$Z_i = W_i + \epsilon$</td>
<td>$Z_i</td>
</tr>
<tr>
<td>CG</td>
<td>$Y_i = \begin{cases} k &amp; \text{if } \beta_{k-1} &lt; W_i \leq \beta_k \ 0 &amp; \text{otherwise} \end{cases}$</td>
<td>$W_i \sim \text{MVN}(\mathbf{X}_B, H(\phi))$</td>
<td>$W_i \sim \text{MVN}(\mathbf{X}_B, H(\phi))$</td>
</tr>
</tbody>
</table>

We compare our results to those of three other models, selected because they are logical models for ordered categorical data that are (or may be) used in practice and can be fit using available software such as R or winBUGS (Lunn et al., 2000). The first is the standard (non-spatial) cumulative probit model, which can be cast in the latent variable framework (Agresti, 2002, pp. 277–283). We also consider two generalized linear spatial (mixed) models (GLSMs). Several authors have described GLSMs for distributions in the exponential family (Diggles et al., 1998; Christensen et al., 2006; Eidsvik et al., 2009, among others).

Table 1 describes and highlights the differences among the four models we consider (the non-spatial model, two GLSMs, and CG model). All four models are based on a cumulative probit framework which links an ordered categorical response $Y$ to a normally distributed latent random variable $Z$ ($W$ for the CG model). The CG model directly links the multivariate Gaussian distribution capturing the spatial correlation to the discrete response $Y$, whereas the other models discretize (conditionally) independent random variables $Z$. The GLSMs are naturally expressed as hierarchical models, where a spatial random effect is incorporated into the standard cumulative probit framework. The two GLSM models differ only in the level of the hierarchy at which $\mathbf{X}_B$ is specified. In fact, GLSM$_1$ can be described as the hierarchically centered version of GLSM$_2$ since the random effect $W$ is now centered about the mean term, $\mathbf{X}_B$. The two models produce different conditional distributions in the Bayesian MCMC algorithm. Hierarchical centering has been shown to improve MCMC convergence for some problems (Gelfand et al., 1995, 1996).

We do not expect the four models to yield the same results, but mainly compare them to point out resulting differences. This is of practical importance since each of these models might be fit for the same purpose of predicting an ordered categorical response or investigating the relationship between the categorical response and a covariate.

The non-spatial cumulative probit model is fit via maximum likelihood within R, using the polr function in the MASS package (Venables and Ripley, 2002). Fitting the Bayesian version of the non-spatial model with priors specified as previously discussed, gives nearly identical results. The CG model is fit in R, using the algorithm and priors discussed in Sections 2 and 3. GLSM$_1$ and GLSM$_2$ are fit in winBUGS using the same priors specified for the CG model.

5.3. Prediction and estimation results

A primary goal of this analysis is to predict the categorical benthic IBI narrative for unobserved locations in Montgomery County. Toward this end, we evaluate predictions for $m = 100$ hold-out locations.

After tuning the algorithm to obtain adequate convergence behavior, the CG and GLSM$_1$ results are based on one run of the MCMC algorithm for 8000 iterations after a burn-in of 2000. The GLSM$_2$ results are only based on 3000 iterations after a burn-in of 2000 because winBUGS repeatedly crashed shortly after 5000 total iterations. Table 2 compares prediction outcomes among the four models, showing the superior prediction of the Clipped Gaussian model over the naive non-spatial cumulative probit model. The Clipped Gaussian model also outperforms both GLSM models, particularly in the extreme categories, as evidenced by the predicted category compositions. This may have important practical implications if, for example, only the sites predicted as poor are to be remediated.

In Table 3, we also compare the predictions using Bayesian expected loss (BEL), as discussed in Section 4. Note the difference between the estimated BEL and validation BEL for each model. A relatively large difference indicates a model is making incorrect predictions accompanied by false confidence (i.e. a relatively large proportion of the total probability is allocated to the mispredicted category). While GLSM$_1$ results in the lowest values for all four BEL measures, it also reveals the largest difference between the estimated and validation measures for both loss functions. A comparison of validation BEL$_1$ to validation BEL$_2$ provides information about how far the mispredictions are from the true value. A larger difference between the values indicates predictions farther from one category away from the true value. GLSM$_1$ and CG performed best in this regard.
In addition to the prediction results, in Table 4 we report and compare the posterior means and 95% posterior intervals for the most interesting parameters, $\beta_1$ and $\phi$. The 95% Wald’s confidence interval for the non-spatial model and the 95% posterior intervals for GLSM and CG exclude zero, indicating the benthic IBI score tends to be lower for sites with higher measures of conductivity, as expected. The posterior means for the spatial models are closer to zero than the estimate from the non-spatial model, indicating that some of the spatial structure is explained through the correlation, rather than solely through the covariate. The estimates of $\beta_1$ are more similar between GLSM and CG than they are between the two GLSM models. This corresponds to the similarities regarding the level at which $X_8$ is specified. However, note the narrower interval excluding zero for the CG model as compared to the GLSM, indicating better accuracy in the estimation which may reflect the omission of the unit variance (essentially a fixed “nugget” effect) coming from the additional stage in the hierarchical specification of the GLM.

The estimate of the correlation scale parameter, $\phi$, may be used to make inferences about the effective (or practical) range of spatial correlation in the latent Gaussian process. The effective range is commonly defined as the distance at which the correlation reaches 0.05, which is calculated as $-\log(0.05)/\phi \approx 3/\phi$ for the CG model and $-\log(0.1)/\phi \approx 2.3/\phi$ for the GSLM models after accounting for the first stage fixed “nugget” effect and $\sigma^2 = 1$ (Banerjee et al., 2004, page 35). From the empirical semi-variogram shown in Fig. 4, we visually estimate the effective range to be about 3.5, before accounting for the spatial trend explained by conductivity. The estimates of $\phi$ from CG and GLSM are both practically reasonable, with most plausible values for the effective range falling between 1.2 and 6.2 for CG, and 0.9 and 2.9 for GLSM. For GLSM, the most plausible values fall between 0.3 and 1.1, noticeably smaller than those from GLSM and CG. Again, we report the results of the three models to emphasize their differences, not because we expect them to be the same. We suggest that more care may need to be taken when implementing such models for ordered categorical data than for analogous hierarchical models for a continuous response. The model comparisons presented here have sparked future work to better describe the differences both analytically and computationally, however this is beyond the scope of the current paper.

### 6. Discussion

The Clipped Gaussian model presented in this paper provides a logical and important extension to the current literature surrounding spatial models for point-referenced ordered categorical data. The application presented 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 on-going 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).

In addition to the application presented in this paper, we have applied the Clipped Gaussian model to a wide variety of simulated data sets (Higgs, 2007). The data sets were generated under the both the Clipped Gaussian model and the GLSM.
model described in Section 5.2 for a random sample of locations selected from a study region. We again assessed prediction using a hold-out set and assessed estimation by comparison of posterior distributions to the data-generating values. In fitting the CG model to over 150 different sets of artificial data, on average 65% of the 50 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. Not surprisingly, less patchy realizations (more smooth) were generally accompanied by better predictions than those from more patchy (less smooth) realizations. Consistent with the application to the benthic IBI data, results indicate the model does a particularly good job predicting values in the extreme categories. Posterior distributions were on average centered near the data-generating value. An important observation from visual examination of the data sets generated under the Clipped Gaussian model is the striking variability among realizations generated under the same parameter values. This is encouraging for two reasons: it speaks to the appropriateness of assuming the Clipped Gaussian model is the data-generating model over a wide range of spatial categorical data sets, and it demonstrates the success of the model at estimation and prediction over a wide range of realizations.

The truncated plurigaussian approach mentioned in Section 1 may hold future promise for extending the ability of the Clipped Gaussian method to successfully model an even larger variety of categorical random fields. In particular, it would allow a relaxation of the ordering of the categories and would perhaps allow the method to be used for categorical data in general. For example, it would theoretically allow any two categories to “touch” in space. Our Clipped Gaussian model, as developed thus far, is appropriate for strictly ordered categories, assuming it is unlikely for non-adjacent categories to be observed next to each other at very small distances. The goal of the plurigaussian approach as described in the geosciences literature is to simulate from a categorical random field in order to understand the variability among realizations (LeLoc’h and Galli, 1997). The models are fit through a series of user controlled steps, rather than under a unified statistical modeling framework.

Another 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., 2009), and non-parametric methods (Kammann and Wand, 2003).

As mentioned in Section 5.3, on-going work involves understanding the theoretical and practical differences between the generalized linear spatial modeling (GLSM) strategy and the Clipped Gaussian strategy. An understanding of practical scenarios for which each model is more appropriate than the other would benefit applied statisticians and researchers. 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, 148). However, the issues for ordered categorical data are more complex than those for continuous data and are intertwined with computational issues. Our results suggest that the Clipped Gaussian model may have better predictive ability, especially when modeling a relatively smooth realization of a categorical random field, or when prediction of the extreme categories is of particular importance.

Acknowledgements

The work was supported by the STAR Research Assistance Agreement CR-829095 awarded by the US 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 authors 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. Hoeting’s research was also supported by NSF Grant DEB-0717367. We would also like to thank three anonymous reviewers for their time and excellent comments used to improve the paper. Finally, we owe thanks to Victor De Oliveira for providing key insights on our work, as well as Kathryn Irvine for her time and helpful suggestions.

Appendix. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.csda.2010.02.024.

References
