Spatial Generalized Linear Models

Joshua Radack and Hyoin An

April 28, 2021

1 Introduction

In many areas of statistical application, an analyst may encounter data which is spatially-indexed. In such scenarios, one may believe that dependence between samples is related to their spatial arrangement and proximity with other sample points. There is a rich literature on spatial (or geostatistical) methods which model data assuming a spatial dependence structure.

Formally, let $\{Y(\boldsymbol{x}) : \boldsymbol{x} \in B\}$ be a stochastic process defined on some spatial region $B \subset \mathbb{R}^p$ (typically, $p = 2$). The spatial dependence is characterized through a positive semi-definite covariance function $\text{Cov}(Y(\boldsymbol{x}), Y(\boldsymbol{x}')) = C_Y(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta})$ for all $\boldsymbol{x}, \boldsymbol{x}' \in B$, where $\boldsymbol{\theta}$ is a vector of parameters. The particular choice for C_Y is largely dictated by the application at hand, but all such functions share the general idea that points closer together in space will be more correlated with one another compared to points farther apart. In practice, we only get to observe $Y(\boldsymbol{x})$ at a finite number of locations, say $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \in B$. Collecting these into an observation vector, $\boldsymbol{Y} = (Y(\boldsymbol{x}_1), ..., Y(\boldsymbol{x}_n))$, we define the mean vector and covariance matrix to be $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{Y})$ and $\Sigma = \text{Cov}(\boldsymbol{Y})$.

Geostatistical methods typically assume that $\{Y(\boldsymbol{x})\}$ is a Gaussian Process, a stochastic process where any finite collection of the random variables has a multivariate normal distribution. However, this normality assumption can be untenable in a variety of situations.

For example, consider a dataset explored in Zhang (2002) and Bonat & Ribeiro (2015) concerning Rhizoctonia, a type of fungal root rot which can hinder the growth of crops such as wheat and barley. The data were collected by sampling 15 barley plants from 100 locations on the Cunningham Farm in Pullman, Washington. At each sampling site, the number of crown roots and the number of infected crown roots of the barley plants were counted. In this scenario, a binomial distribution seems more appropriate than a normal distribution for the data. Such datasets motivate the development of spatial generalized linear models to extend geostatistical methods to handle non-Gaussian spatially dependent data.

In this report, we will explore several methods proposed for the modeling of spatial data in a non-Gaussian context. In section 2 we discuss both a frequentist approach (generalized estimating equations) and Bayesian approach (generalized linear mixed models) to spatial GLMs. In section 3 we fit one of the proposed models to a real-world dataset to both estimate parameters and make predictions at unobserved locations. Finally section 4 ends with a summary of the report.

2 Methods

2.1 Frequentist approach

As noted in section 1, suppose $\boldsymbol{Y}$ is an $n \times 1$ vector of spatially-dependent observations with mean $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$ and variance-covariance matrix $\text{Cov}(\boldsymbol{Y}) = \Sigma$, which is a positive-definite matrix whose elements are a function of $\boldsymbol{\mu}$. Additionally, assume a link function g relating the mean to a linear combination of covariates, $\boldsymbol{\eta} = g(\boldsymbol{\mu}) = D\boldsymbol{\beta}$, and a variance function $v(\mu_i)$ where $\text{Var}(Y_i) = a(\phi_i)v(\mu_i)$.

Gotway and Stroup (1997) introduced a method for modeling spatially dependent non-Gaussian data using a quasi-likelihood framework. Mimicking the form of the quasi-score function for independent data, they proposed the quasi-likelihood Q should be governed by

the relation

$$\frac{\partial Q(\boldsymbol{\mu}; \boldsymbol{y})}{\partial \boldsymbol{\mu}} = \Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu}),$$

which is the quasi-score for dependent data. Differentiating Q with respect to $\boldsymbol{\beta}$ yields the score function

$$\boldsymbol{U} = \Delta^T \Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu}),$$

where Δ is an $n \times p$ matrix with i-j element $\dfrac{\partial \mu_i}{\partial \beta_j}$. A consistent estimator for $\boldsymbol{\beta}$ can be obtained as a solution to $\boldsymbol{U} = \boldsymbol{0}$ if the covariance matrix can be written as

$$\Sigma = v_{\boldsymbol{\mu}}^{1/2} R(\boldsymbol{\theta}) v_{\boldsymbol{\mu}}^{1/2},$$

where $v_{\boldsymbol{\mu}} = \text{diag}(v(\mu_i))$ is the diagonal matrix of variance functions $v(\mu_i)$ and R is the correlation matrix for $\boldsymbol{Y}$ parameterized by $\boldsymbol{\theta}$.

To estimate $\boldsymbol{\beta}$, we recognize that $\Delta = MD$, where $M = \text{diag}(\partial \mu_i / \partial \eta_i)$. Substituting this into the equation for $\boldsymbol{U}$ and applying the method of scoring, we get

$$D^T W D \boldsymbol{\beta} = D^T W \boldsymbol{y}^*,$$

where $W = M^T \Sigma^{-1} M$ and $\boldsymbol{y}^* = \boldsymbol{\eta} + M^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$. This mirrors the estimating equation in the independent data setting, and iterated reweighted least squares can be used to compute the estimate for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_G$.

Oftentimes, a key task in geostatistical applications is to predict the process values at unobserved spatial locations. To this end, let $\boldsymbol{Y}^* = (Y(\boldsymbol{x}_1^*), ..., Y(\boldsymbol{x}_k^*))^T$ be the random variables of the process at k unobserved spatial locations $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, ..., \boldsymbol{x}_k^*)^T$. We define the mean vectors and covariance matrix for these two random vectors,

$$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu}(\boldsymbol{x}),\ \mathbb{E}(\boldsymbol{Y}^*) = \boldsymbol{\mu}(\boldsymbol{x}^*)$$

$$\text{Cov}\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{Y}^* \end{pmatrix} = \begin{pmatrix} \Sigma_{nn} & \Sigma_{nk} \\ \Sigma_{kn} & \Sigma_{kk} \end{pmatrix}.$$

When $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$ are normally distributed and the mean functions can be written as linear functions of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\mu}(\boldsymbol{x}) = X\boldsymbol{\beta}$ and $\boldsymbol{\mu}(\boldsymbol{x}^*) = X^*\boldsymbol{\beta}$, then the best heterogeneous linear predictor has the form

$$p(\boldsymbol{Y}^*|\boldsymbol{Y}) = \Sigma_{kn}\Sigma_{nn}^{-1}\boldsymbol{Y} + (D^* - \Sigma_{kn}\Sigma_{nn}^{-1}D)\boldsymbol{\beta}.$$

This is known as the simple kriging predictor in geostatistics, and if $\boldsymbol{\beta}$ is replaced by its generalized least squares estimate, then the we have the best linear unbiased predictor (BLUP), also known as the universal kriging predictor. To generalize this to non-Gaussian data where $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, first we rewrite the kriging predictor as

$$p(\boldsymbol{Y}^*|\boldsymbol{Y}) = D^*\hat{\boldsymbol{\beta}}_G + \Sigma_{kn}\Sigma_{nn}^{-1}(\boldsymbol{Y} - D\hat{\boldsymbol{\beta}}_G) = \hat{\boldsymbol{\mu}}(\boldsymbol{x}^*) + \Sigma_{kn}\Sigma_{nn}^{-1}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}(\boldsymbol{x})).$$

Then, using the inverse link function h, we can get $\hat{\boldsymbol{\mu}}(\boldsymbol{x}) = h(D\hat{\boldsymbol{\beta}}_G)$ and $\hat{\boldsymbol{\mu}}(\boldsymbol{x}^*) = h(D^*\hat{\boldsymbol{\beta}}_G)$ and plug these into the kriging predictor equation.

2.2 Bayesian approach

There is a stream of literature that considers models in a Bayesian framework for data with non-Gaussian responses that are spatially correlated. Diggle, Tawn, and Moyeed (1998) especially focused on Poisson and binomial responses and suggested a spatial generalized linear mixed model that extends classical geostatistical methodologies. Let $S(\boldsymbol{x})$ be the underlying spatial surface where the responses vary, then $\hat{S}(\boldsymbol{x})$ is a so-called kriging (spatially interpolating) predictor.

Consider the general random effects model

$$\eta_i = g(\mu_i) = \boldsymbol{d}_i^T \boldsymbol{\beta} + U_i, \quad i = 1, ..., n$$

for some known link function $g(\cdot)$, $\mu_i = \mu(\boldsymbol{x}_i) = \mathbb{E}[Y_i|U_i]$, covariates $\boldsymbol{d}_i = \boldsymbol{d}(\boldsymbol{x}_i)$, unknown parameters $\boldsymbol{\beta}$, and the random effects U_i which could be dependent on each other. The natural extension of this model with spatial random effects would be then to let U_i be the spatial process $S(\boldsymbol{x}_i)$ and obtain

$$\mathbb{E}[Y_i|S(\boldsymbol{x}_i)] = \mu(\boldsymbol{x}_i), \quad \text{where} \quad g(\mu(\boldsymbol{x}_i)) = \boldsymbol{d}(\boldsymbol{x}_i)^T \boldsymbol{\beta} + S(\boldsymbol{x}_i) \tag{1}$$

for $i = 1, ..., n$, which is in align with the classical generalized linear model formulation. Due to the random spatial effects $S(\boldsymbol{x})$, (1) is called a spatial generalized linear mixed model, proposed by Diggle et al. (1998). They assume the followings to be satisfied.

(a) S is a stationary Gaussian process with $\mathbb{E}[S(\boldsymbol{x})] = 0$ and

$$C_{\boldsymbol{\theta}}(S(\boldsymbol{x}), S(\boldsymbol{x}')) = \mathrm{Cov}[S(\boldsymbol{x}), S(\boldsymbol{x}')] = \sigma^2 \rho(\boldsymbol{x} - \boldsymbol{x}'),$$

where $\boldsymbol{\theta}$ is the set of parameters consisting of σ^2 and any parameters related to the correlation structure of S.

(b) $Y_1, ..., Y_n$ are conditionally independent given S with distributions

$$f_i(y|S(\boldsymbol{x}_i)) = f(y; \mu_i),$$

where $\mu_i = \mathbb{E}[Y_i|S(\boldsymbol{x}_i)]$, $i = 1, ..., n$.

(c) $g(\mu_i) = S(\boldsymbol{x}_i) + \boldsymbol{d}_i^T \boldsymbol{\beta}$, for some known link function $g(\cdot)$, explanatory variables $\boldsymbol{d}_i = \boldsymbol{d}(\boldsymbol{x}_i)$ and parameters $\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ denotes all the regression parameters.

The kriging predictor $\hat{S}(\boldsymbol{x}) = \mathbb{E}[S(\boldsymbol{x})|\boldsymbol{Y}]$ under these assumptions is called the generalized linear predictor for $S(\boldsymbol{x})$.

Let $\boldsymbol{S}$ and $\boldsymbol{S}^*$ denote the set of values of $S(\boldsymbol{x})$ at the sampling locations $\boldsymbol{x}_i, i = 1, ..., n$ and at the locations $\boldsymbol{x}_j^*, j = 1, ..., m$ for which predictions are required, respectively. In addition, let $g_k(\boldsymbol{s})$ denote the multivariate normal probability density function of the first k elements of $(\boldsymbol{S}, \boldsymbol{S^*})$, whose ith element is denoted by s_i, $i = 1, ..., n + m$. Under assumptions (a)-(c), the unconditional density of $\boldsymbol{Y}$ is then

$$f(\boldsymbol{y}) = \int \left\{ \prod_{i=1}^{n} f_i(y_i|s_i) \right\} g_n(\boldsymbol{s}) d\boldsymbol{s}$$

and the unconditional joint density of $(\boldsymbol{S^*}, \boldsymbol{Y})$ is

$$\int \left\{ \prod_{i=1}^{n} f_i(y_i|s_i) \right\} g_{n+m}(\boldsymbol{s}) d\boldsymbol{s}.$$

Then, the generalized linear predictor for location $\boldsymbol{x}_j^*$ is

$$\hat{S}(\boldsymbol{x}_j^*) = \mathbb{E}[S(\boldsymbol{x}_j^*)|\boldsymbol{Y}] = \frac{\mathbb{E}_{1,j}}{f(\boldsymbol{y})}$$

with the associated prediction variance

$$\nu(\boldsymbol{x}_j^*) = \mathbb{E}[S(\boldsymbol{x}_j^*)^2|\boldsymbol{Y}] - \mathbb{E}[S(\boldsymbol{x}_j^*)|\boldsymbol{Y}]^2 = \frac{\mathbb{E}_{2,j}}{f(\boldsymbol{y})} - \hat{S}(\boldsymbol{x}_j^*)^2$$

where

$$\mathbb{E}_{r,j} \equiv \int s_{n+j}^r \left\{ \prod_{i=1}^{n} f_i(y_i|s_i) \right\} g_{n+m}(\boldsymbol{s}) d\boldsymbol{s} ds_{n+j}$$

for some $r > 0$ and $j = 1, ..., m$.

Calculating $f(\boldsymbol{y}), \mathbb{E}_{1,j}, \mathbb{E}_{2,j}$ involves intractable integration, especially when n is large. To tackle this problem, Diggle et al. (1998) construct a Markov Chain Monte Carlo (MCMC) algorithm to obtain random samples from the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{S}, \boldsymbol{\beta})|\boldsymbol{Y}$ for infer-

ence on $(\boldsymbol{\theta}, \boldsymbol{S}, \boldsymbol{\beta})$ as well as from the posterior distribution of $\boldsymbol{S}^*|(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{S}, \boldsymbol{\beta})$ for prediction at unobserved locations. By assumption (b), notice that

$$p(\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{\beta}) = \prod_{j=1}^{n} f(y_i|s_j, \boldsymbol{\beta}),$$

where $f_i(y|S(\boldsymbol{x}_i)) \equiv f(y; \mu_i)$. With this, we can write the conditional posterior distributions for each of $(\boldsymbol{\theta}, \boldsymbol{S}, \boldsymbol{\beta})$ as follows.

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{\beta}) &= \pi(\boldsymbol{\theta}|\boldsymbol{S}) \propto p(\boldsymbol{S}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
\pi(S_j|\boldsymbol{S}_{-j}, \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\beta}) &\propto p(\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{\beta})p(S_j|\boldsymbol{S}_{-j}, \boldsymbol{\theta}) = \left\{ \prod_{i=1}^{n} f(y_i|s_i, \boldsymbol{\beta}) \right\} p(S_j|\boldsymbol{S}_{-j}, \boldsymbol{\theta}) \\
\pi(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{\theta}) &= \pi(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{S}) \propto p(\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{\beta})p(\boldsymbol{\beta}) = \left\{ \prod_{j=1}^{n} f(y_i|s_j, \boldsymbol{\beta}) \right\} p(\boldsymbol{\beta})
\end{aligned}$$

where $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\beta})$ are prior distributions. Diggle et al. (1998) choose the uniform prior for $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\beta})$. This might be a convenient choice of prior, but in case we want the priors to be non-informative we can consider Jeffrey's prior or Reference prior, which are not necessarily uniform. Also, if we have any available information apriori, then we might want to include it in a prior distribution. For example, the spatial covariance parameters tend to be correlated on each other, and we can try to put a joint prior on $\boldsymbol{\theta}$ containing the proper correlation. Christensen and Waagepetersen (2002) consider Gaussian distribution for $\boldsymbol{\beta}$ after reparametrization for better mixing Markov chains.

By assumption (a), we know $p(\boldsymbol{S}|\boldsymbol{\theta})$ follows a multivariate normal distribution, which implies that S_i follows a univariate conditional normal posterior distribution , i.e. $\pi(S_j|\boldsymbol{S}_{-j}, \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\beta})$ is a conditional normal density for $j = 1, ..., n$. The MCMC algorithm Using the Metropolis-Hastings step (Smith and Roberts, 1993) is as follows.

Step 1 (Update $\boldsymbol{\theta}$) Generate $\boldsymbol{\theta}' \sim p(\boldsymbol{\theta})$ and accept $\boldsymbol{\theta}'$ with probability

$$\Delta(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min \left\{ \frac{p(\boldsymbol{S}|\boldsymbol{\theta}')}{p(\boldsymbol{S}|\boldsymbol{\theta})}, 1 \right\}$$

Step 2 (Update $\boldsymbol{S}$) Generate $\boldsymbol{S}' = (S_1', ..., S_n'), S_j' \sim q(S_j, S_j')$, where $q(S_j, S_j') = p(S_j'|\boldsymbol{S}_{-j}, \boldsymbol{\theta})$ and accept S_j' for all $j = 1, ..., n$ with probability

$$\Delta(S_j, S_j') = \min\left\{\frac{f(y_j|s_j', \boldsymbol{\beta})}{f(y_j|s_j, \boldsymbol{\beta})}, 1\right\}$$

Step 3 (Update $\boldsymbol{\beta}$) Generate $\boldsymbol{\beta}' \sim q(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ and accept $\boldsymbol{\beta}'$ with probability

$$\Delta(\boldsymbol{\beta}, \boldsymbol{\beta}') = \min\left\{\frac{\prod_{j=1}^n f(y_j|s_j, \boldsymbol{\beta}')q(\boldsymbol{\beta}', \boldsymbol{\beta})}{\prod_{j=1}^n f(y_j|s_j, \boldsymbol{\beta})q(\boldsymbol{\beta}, \boldsymbol{\beta}')}, 1\right\}$$

Step 4 (Sample $\boldsymbol{S}^*$) Generate $\boldsymbol{S}^*|(\boldsymbol{S}, \boldsymbol{\theta}) \sim MVN(\Sigma_{12}^T\Sigma_{11}^{-1}\boldsymbol{S},\ \Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})$, where $\Sigma_{11} = \mathrm{Var}(\boldsymbol{S}), \Sigma_{12}\,\mathrm{Cov}(\boldsymbol{S}, \boldsymbol{S}^*), \Sigma_{22} = \mathrm{Var}(\boldsymbol{S}^*)$.

There are other algorithms developed in the literature under the spatial GLMM setup, and we introduce two of them here. Christensen and Waagepetersen (2002) point out that the update of S_j in Diggle et al. (1998) is computationally demanding due to the matrix inversion in calculation of conditional mean and variance. They propose replacing the Metropolis-Hastings step with Langevin-Hastings updates in the context of Poisson log-normal data, using the truncated Langevin-Hastings proposal distribution, which allows for simultaneous update for the random effects and leads to more efficient sampling.

Instead of using the full MCMC sampling scheme, Zhang (2002) suggests combining the Monte Carlo method with an EM gradient algorithm. They treat spatial random effects as missing data and impute these missing data, using the EM algorithm, and Monte Carlo samples are used to calculate the approximate conditional expectation in the E-step in the EM gradient algorithm. One advantage of their algorithm is that it provides the maximum likelihood estimates of parameters and the minimum mean-squared error estimates of the spatial effect at the observed locations. Also, the computation could be more efficient by avoiding the MCMC simulation.

2.3 A more recent method

The simulation-based algorithms, such as MCMC, require a tuning process and convergence analysis. This poses a challenge as the number of locations sampled increases, since the computational costs for MCMC algorithms will increase for two reasons. The first is caused by having to invert the covariance matrix of latent Gaussian process $\boldsymbol{S}$, and with changing covariance parameters each step of the MCMC, this inversion must happen frequently. The second is the sheer number of random spatial effects S_i, $i = 1, ..., n$, which must be updated. Sampling these with a Metropolis-Hastings update can lead to convergence issues as the space is not explored efficiently.

Instead of relying on the MCMC algorithm, one of the more recent methods for fitting spatial generalized linear mixed models in a likelihood analysis framework is proposed by Bonat and Ribeiro (2015). Their method is based on the Laplace approximation (Tierney and Kadane, 1986) and provides the maximized log-likelihood value, allowing for model selection and tests.

The Laplace approximation is utilized to calculate the intractable integral function of the marginal log-likelihood of the parameters, which is maximized to estimate the parameters in the model. The use of the Laplace approximation can be effective, especially when the computation of the MCMC algorithm is expected to be complex. Bonat and Ribeiro (2015) showed in a simulation study that their method with Laplace approximation had adequate results in parameter estimation under a couple of standard applications. According to Tierney and Kadane (1986), the accuracy of the approximation is of order $O(n^2)$, implying that the approximation tends to work well for large sample. With more sophisticated problems, however, this might not always be the case. If the parameters are correlated, the profile likelihood would be unstable and not be reliable. The parameters in spatial covariance, for example, often are heavily correlated. Therefore, the implementation of the Laplace approximation needs to be conducted and interpreted carefully.

For the model construction, Bonat and Ribeiro (2015) use the same assumptions as in

section 2.2, but they specify a Gaussian spatial process as

$$S(\boldsymbol{x}) = \sigma \boldsymbol{U}(\boldsymbol{x}; \phi) + \tau \boldsymbol{Z},$$

where $\boldsymbol{U}(\boldsymbol{x}; \phi)$ denotes spatially correlated random effects (a unit variance Gaussian random field) with a correlation function $\rho(u, \phi)$ and $\tau \boldsymbol{Z} \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{I})$.

Their idea is as follows. Write $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \phi, \psi)$ for all the parameters in the model, where ϕ is a parameter that decides correlation function $\rho(\cdot, \phi)$ and ψ is a parameter that decides $f(\cdot; \psi)$. Instead of using simulation methods, the goal is then to estimate $\boldsymbol{\alpha}$ by maximizing the marginal likelihood function $L(\boldsymbol{\alpha})$,

$$L(\boldsymbol{\alpha}; \boldsymbol{y}(\boldsymbol{x})) = \int_{\mathbb{R}^n} f(y(\boldsymbol{x})|S(\boldsymbol{x})) f(S(\boldsymbol{x})) dS(\boldsymbol{x}).$$

To approximate this integral, consider the following Laplace approximation (Tierney & Kadane, 1986),

$$\int_{\mathbb{R}^n} \exp\{Q(\boldsymbol{u})\} d\boldsymbol{u} \approx (2\pi)^{n/2} |-Q''(\hat{\boldsymbol{u}})|^{-1/2} \exp\{Q(\hat{\boldsymbol{u}})\}, \tag{2}$$

where $Q(\boldsymbol{u})$ is a known, uni-modal, and bounded function, $\hat{\boldsymbol{u}}$ is the value for which $Q(\boldsymbol{u})$ is maximized, and $Q''(\hat{\boldsymbol{u}})$ is called Hessian.

For simplicity, if we assume that $f(\boldsymbol{y}(\boldsymbol{x})|S(\boldsymbol{x}))$ is an exponential family, i.e.

$$f(\boldsymbol{y}(\boldsymbol{x})|S(\boldsymbol{x})) = \exp\left\{\boldsymbol{y}(\boldsymbol{x})^T (\boldsymbol{D}\boldsymbol{\beta} + S(\boldsymbol{x})) - 1^T b(\boldsymbol{D}\boldsymbol{\beta} + S(\boldsymbol{x})) + 1^T c(\boldsymbol{y}(\boldsymbol{x}))\right\}$$

for some known functions $b(\cdot)$ and $c(\cdot)$. Note that this is in the canonical form, leading to the simple functional form of the $Q(\cdot)$ for $S(\boldsymbol{x})$ which is shown on the next page.

We can write the multivariate Gaussian density function of $S(\boldsymbol{x})$ as follows,

$$f(S(\boldsymbol{x}); V) = (2\pi)^{-n/2}|V|^{-1/2} \exp\left\{-\frac{1}{2}S(\boldsymbol{x})^T V^{-1} S(\boldsymbol{x})\right\},$$

and then the likelihood function can be expressed as

$$L(\boldsymbol{\alpha}; \boldsymbol{y}(\boldsymbol{x})) = \int_{\mathbb{R}^n} \exp\{Q(S(\boldsymbol{x}))\} dS(\boldsymbol{x}),$$

where

$$\begin{aligned} Q(S(\boldsymbol{x})) =& \boldsymbol{y}(\boldsymbol{x})^T(\boldsymbol{D}\boldsymbol{\beta} + S(\boldsymbol{x})) - 1^T b(\boldsymbol{D}\boldsymbol{\beta} + S(\boldsymbol{x})) + 1^T c(\boldsymbol{y}(\boldsymbol{x})) \\ & - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log|V| - \frac{1}{2}S(\boldsymbol{x})^T V^{-1} S(\boldsymbol{x}). \end{aligned}$$

Then, by finding $\hat{\boldsymbol{s}} = \max Q(S(\boldsymbol{x}))$, we can approximate $L(\boldsymbol{\alpha}; \boldsymbol{y}(\boldsymbol{x}))$ using the equation (2) without calculating the integration. We can solve the maximization problem using various numerical methods in practice, such as Newton-Raphson methods, with the number of iterations usually being small (Tierney & Kadane, 1986). By taking the log to the approximation of the likelihood, we then can obtain the log-likelihood and the maximum likelihood estimator of $\boldsymbol{\alpha}$ that maximizes the log-likelihood.

3 Application

Recall the Rhizoctonia dataset from section 1 (Zhang, 2002; Bonat & Ribeiro, 2015). Samples at each location consisted of the number of crown roots counted and the number of infected crown roots counted on the extracted barley plants, ranging from $(80, 197)$ and $(1, 70)$, respectively. Since the incidence rate can vary across space, mapping the incidence rate across the entire farm could provide valuable information for the farmer wishing to target a treatment for the rot to specific locations of higher incidence.

A map of the observed proportions of root rot are found in figure 1. The units along

each axis is in meters, with the region having an area of about 36-ha. Sampling locations were chosen so that each site was at least 30-100 meters apart from any other site (Paulitz, et al. 2003). In the plot, the radius of each point is scaled to the observed proportions, which ranged from 0.0079 to 0.483. From this map, we already notice the presence of spatial correlation of the observed proportions, with points of similar magnitude clustered together.

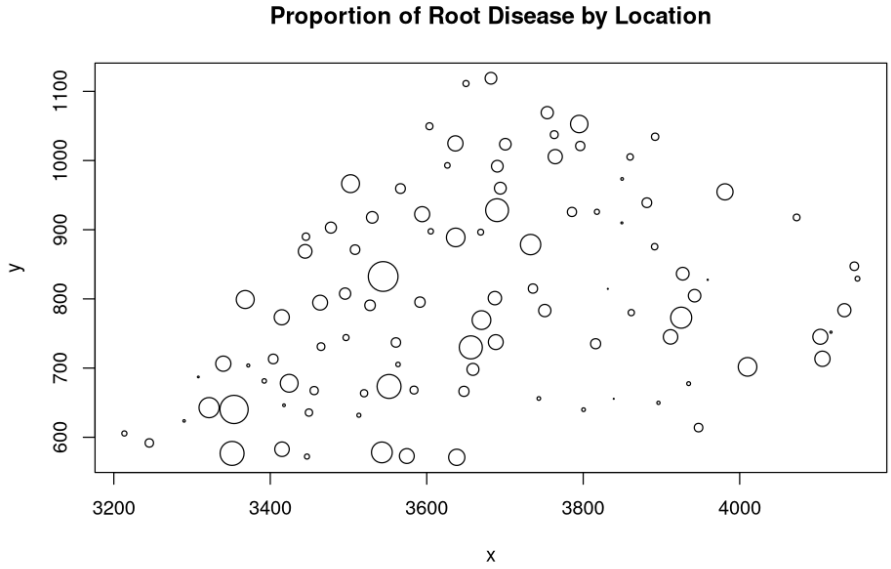


Figure 1: Map of observed proportion of root rot

In light of this, we consider the following GLMM structure for the process: let $\{Y(\boldsymbol{x}) : \boldsymbol{x} \in B\}$ be spatial process of root rot count, where

$$Y(\boldsymbol{x})|S(\boldsymbol{x}) \sim \text{Binomial}(m_{\boldsymbol{x}}, p_{\boldsymbol{x}}) \quad \eta(\boldsymbol{x}) = g(\mu(\boldsymbol{x})) = \beta_0 + S(\boldsymbol{x}).$$

We assume $S(\boldsymbol{x})$ is a stationary mean-zero Gaussian process with spherical covariance function

$$C_S(t) = \begin{cases} \tau^2 \mathbf{1}(t=0) + \sigma^2 \left[1 - \frac{3}{2}\frac{t}{\phi} + \frac{1}{2}\left(\frac{t}{\phi}\right)^3\right], & 0 \le t \le \phi \\ 0, & t > \phi \end{cases}$$

for $t = \boldsymbol{x} - \boldsymbol{x}'$ for any $\boldsymbol{x}, \boldsymbol{x}' \in B$. This covariance function was suggested by Zhang (2002) through analyzing the empirical variogram and noting it was flat after a certain distance. The spherical covariance function also has the advantage of being compact, making covariance matrix inversions more computationally efficient. The link function g is the canonical logit link and we assume an intercept-only fixed effect. We assume the $Y(\boldsymbol{x})$ are independent conditional on $S(\boldsymbol{x})$.

We utilize code made available by Bonat and Ribeiro (2015) to estimate parameters using the Laplace approximation. The parameter estimates and associated standard errrs are found in table 1. The estimates closely match the results obtained through alternate estimation methods such as MCMC and Monte Carlo EM (Zhang, 2002). We notice the standard error for ϕ being extremely large - Bonat and Reibeiro comment that since the sample size is relatively small ($n = 100$), inference based on these standard errors would be unreliable. The fact that covariance parameters tend to be highly correlated with each other only complicates this matter further.

	Estimate	Std. Error
β_0	-1.7183	0.097421
σ^2	0.10571	1.2257
ϕ	148.66	1.1187×10^{64}
τ^2	0.46676	0.43448

Table 1: Parameter Estimates and Standard Errors.

In addition to estimation, we will perform prediction at unobserved locations to form an interpolated map of the binomial probability of infection, $p_{\boldsymbol{x}}$. The reason we choose to perform prediction for this parameter rather than the original process $\{Y(\boldsymbol{x})\}$ itself is because it would necessitate also predicting the binomial count, $m_{\boldsymbol{x}}$, at each prediction location.

Let $\boldsymbol{\eta}^* = (\eta(\boldsymbol{x}_1^*), ..., \eta(\boldsymbol{x}_k^*))$ be the random variables of the link-scale mean at k unobserved spatial locations $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, ..., \boldsymbol{x}_k^*)$ we wish to predict. We have

$$\mathbb{E}(\boldsymbol{\eta}) = \beta_0 \mathbf{1}_n, \ \mathbb{E}(\boldsymbol{\eta}^*) = \beta_0 \mathbf{1}_k$$

$$\text{Cov}\begin{pmatrix}\boldsymbol{\eta}\\ \boldsymbol{\eta}^*\end{pmatrix} = \begin{pmatrix}\Sigma_{nn} & \Sigma_{nk}\\ \Sigma_{kn} & \Sigma_{kk}\end{pmatrix}.$$

Since the $\eta(\boldsymbol{x})$'s are normally distributed, the best heterogeneous linear predictor (known as the universal kriging predictor) has the form of the conditional mean

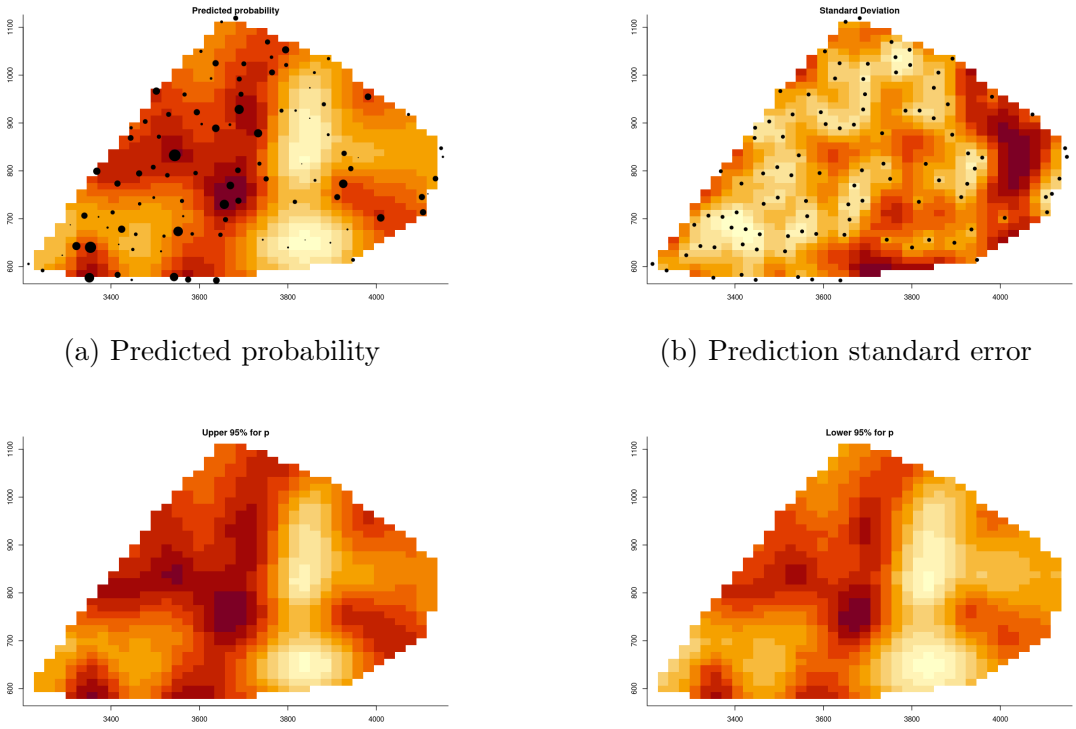
$$p(\boldsymbol{\eta}^*|\boldsymbol{\eta}) = \mathbb{E}(\boldsymbol{\eta}^*) + \text{Cov}(\boldsymbol{\eta}^*, \boldsymbol{\eta})\,\text{Cov}(\boldsymbol{\eta}, \boldsymbol{\eta})^{-1}(\boldsymbol{\eta} - \mathbb{E}(\boldsymbol{\eta}))$$

Once parameter estimates are obtained, we use a universal kriging predictor on the link scale to make predictions for $g(p(\boldsymbol{x}))$ on a grid of locations $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_k^* \in B$.

$$\tilde{\boldsymbol{\eta}}(\boldsymbol{x}^*) = \hat{\beta}_0\mathbf{1}_k + \hat{\Sigma}_{kn}\hat{\Sigma}_{nn}^{-1}(g(\hat{\boldsymbol{p}}) - \hat{\beta}_0\mathbf{1}_n)$$

where the i-th entry of $\hat{\boldsymbol{p}}$ is Y_i/m_i and the covariance matrices are estimated by plugging in the parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\tau}^2, \hat{\sigma}^2, \hat{\phi})$. Standard errors for the predictions can be derived from the estimated prediction variance $\hat{\Sigma}_{kk} - \hat{\Sigma}_{kn}\hat{\Sigma}_{nn}^{-1}\hat{\Sigma}_{nk}$, which are estimated by plugging in the estimated covariance parameters. We then apply the inverse logit link to get predictions on the probability scale, $p(\boldsymbol{x})$.

A panel of four maps in figure 2 shows the results from predicting the probability of root rot across the farm. In figure 2a we see the predictions as a heat map (darker color indicates higher predicted probability) with the observed proportion of root rot overlaid as points (larger radius indicates higher proportion). This provides a smoothed map of the predicted probabilities, indicating areas of higher and lower predicted probability. We note a distinctive ridge line from north to south around the center of the map, which may be connected to some underlying feature of the farm. In figure 2b, we see the the prediction standard errors are lower (lighter heat map color) near the observed locations with a region of higher standard error on the eastern edge of the map. We can also form 95% confidence bounds for the prediction by transforming a 95% interval for the $\boldsymbol{\eta}$'s and using the inverse-



(a) Predicted probability

(b) Prediction standard error

(c) Upper 95% bound for prediction

(d) Lower 95% bound for prediction

Figure 2: Prediction maps

logit function, shown in figures 2c and 2d.

4 Summary

In this report, we have introduced several methods for modeling non-Gaussian data that is spatially indexed. In particular, Gotway and Stroup (1997) extended the quasi-likelihood methods and provided the base for the development of generalized linear models for spatially correlated data. On the other hand, Diggle et al. (1998) proposed a spatial generalized linear mixed model and took a Bayesian approach to generate the posterior samples of the parameters and spatial random effects. In the simulation study, we adopted Bonat and Ribeiro's suggestion (2015) to utilize the Laplace approximation for spatial generalized linear mixed models and confirmed that the model provided the reasonable results with the estimated spatial effects in the non-Gaussian data.

References

[1] Bonat, W. H., & Ribeiro, P. J. (2015) Practical likelihood analysis for spatial generalized linear mixed models. *Environmentrics*, 27(2), 83-89.

[2] Christensen, O. F., & Waagepetersen, R. (2002) Bayesian Prediction of Spatial Count Data Using GLMM. *Biometrics*, 58(2), 280-286.

[3] Diggle, P. J., Tawn, J. A., & Moyeed, R. A. (1998) Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3), 299-350.

[4] 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(2), 157-178.

[5] Paulitz, T.C., Zhang, H., & Cook, R.J. (2003) Spatial distribution of Rhizoctonia oryzae and rhizoctonia root rot in direct-seeded cereals. *Canadian Journal of Plant Pathology*, 25(3), 295-303.

[6] Smith, A. F. M. & Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society*, B(55), 3-23.

[7] Tierney, L., & Kadane, J. B. (1986) Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of American Statistical Association*, 81(393), 82-86.

[8] Zhang, H. (2002) On Estimation and Prediction for Spatial Generalized Linear Mixed Models. *Biometrics*, 58(1), 129-136.