BIOMETRICS 58, 129–136
March 2002

On Estimation and Prediction for Spatial Generalized Linear Mixed Models

Hao Zhang

Program in Statistics, Washington State University, Pullman, Washington 99164-3144, U.S.A.
email: zhanghao@wsu.edu

SUMMARY. We use spatial generalized linear mixed models (GLMM) to model non-Gaussian spatial variables that are observed at sampling locations in a continuous area. In many applications, prediction of random effects in a spatial GLMM is of great practical interest. We show that the minimum mean-squared error (MMSE) prediction can be done in a linear fashion in spatial GLMMs analogous to linear kriging. We develop a Monte Carlo version of the EM gradient algorithm for maximum likelihood estimation of model parameters. A by-product of this approach is that it also produces the MMSE estimates for the realized random effects at the sampled sites. This method is illustrated through a simulation study and is also applied to a real data set on plant root diseases to obtain a map of disease severity that can facilitate the practice of precision agriculture.

KEY WORDS: EM algorithm; Generalized linear mixed models; Metropolis–Hastings algorithm; Spatial interpolation; Variogram.

1. Introduction

Spatial non-Gaussian data, especially count data, arise in many situations in epidemiology, ecology, and agriculture, to name a few. A typical example is the incidence rates for which there are two distinguishing cases: data observed on contiguous administrative regions such as counties and data observed at point locations within a continuous area. The former case arises in disease mapping problems in epidemiology and has been studied by many people (cf., Besag, York, and Mollie, 1991; Waller et al., 1997; and the special issue of *Statistics in Medicine,* 2000, pp. 2201–2593). This article concerns the latter case, where interpolation is needed to predict values at unsampled sites.

We consider a motivating example that consists of spatial non-Gaussian data of *Rhizoctonia* root rot collected on the Cunningham Farm. Located 7 miles north of Pullman, Washington, the 90-acre farm has been direct seeded (i.e., seeded without plowing) to wheat and barley since 1997. One of the major limiting factors to direct-seeded wheat and barley is the root disease *Rhizoctonia* root rot caused by the fungi *Rhizoctonia solani* and *Rhizoctonia oryzae* (Cook and Haglund, 1991; Cook, 1992). These fungi attach to the root system and reduce the ability of plants to take up adequate water and nutrients. The severity of root rot varies in a farm, and a map of severity of the root rot is invaluable in precision agriculture that utilizes site-specific information when applying fungicides, pesticides, or fertilizers. Dr R. J. Cook of Washington State University collected *Rhizoctonia* root rot data in the summer of 2000 at 100 randomly selected sites on the farm. At each sampling site, 15 plants of barley were pulled from the ground and the number of crown roots and infected crown roots were counted for each plant. Then the incidence rate of root rot at each site was obtained as the corresponding ratio. Although the number of crown roots sampled at each site ranged from 80 to 197, the incidence rates are quite skewed and hence non-Gaussian (Figure 1). Even though some transformations might make the data normal, it is unlikely that the transformed data are stationary due to the heterogeneity of sample sizes. On the other hand, it is reasonable to assume that the incidence rate is binomial at each site, with a varying binomial parameter.

Diggle, Tawn, and Moyeed (1998) employed spatial generalized linear mixed models (GLMMs) for spatially dependent non-Gaussian variables observed in a continuous region and considered the minimum mean-squared error (MMSE) prediction under the Bayesian framework. In the present article, we will also use a spatial GLMM to model spatial non-Gaussian data. For any spatial location $\boldsymbol{s}$, let $Y(\boldsymbol{s})$ denote the response variable and $x_1(\boldsymbol{s}), x_2(\boldsymbol{s}), \ldots, x_p(\boldsymbol{s})$ the p observable explanatory variables and let $\{b(\boldsymbol{s}), \boldsymbol{s} \in R^2\}$ be an unobservable spatial random process such that $b(\boldsymbol{s})$ represents the random effect at site $\boldsymbol{s}$ of unknown or unobservable causes unaccounted for by the explanatory variables. The model is defined as follows:

(a) $\{b(\boldsymbol{s}), \boldsymbol{s} \in R^2\}$ is a Gaussian stationary process with $\mathrm{E}b(\boldsymbol{s}) = 0$ for all $\boldsymbol{s}$ and $\mathrm{cov}(b(\boldsymbol{s}+\boldsymbol{h}), b(\boldsymbol{s})) = C(\boldsymbol{h})$ for all $\boldsymbol{s}, \boldsymbol{h} \in \mathbb{R}^2$, where the covariogram $C(\cdot)$ depends on a vector of parameters $\theta \in R^v$.

(b) Conditionally on $\{b(\boldsymbol{s}), \boldsymbol{s} \in R^2\}$, $\{Y(\boldsymbol{s}), \boldsymbol{s} \in R^2\}$ is an independent process and the distribution of $Y(\boldsymbol{s})$ is specified by the conditional mean $\mathrm{E}\{Y(\boldsymbol{s}) \mid b(\boldsymbol{s})\}$.

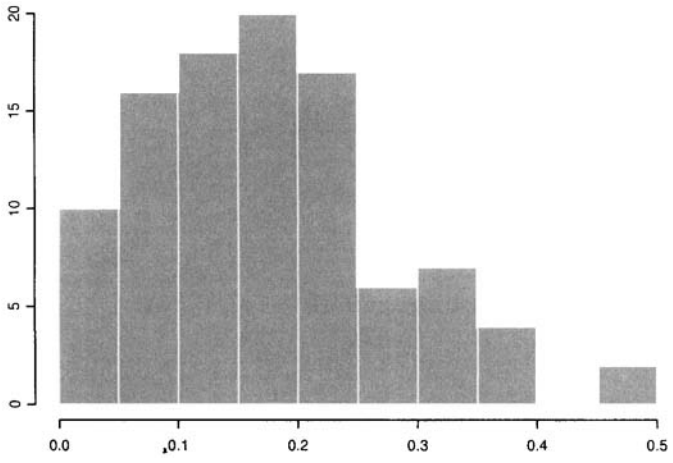

Figure 1. Histogram of incidence rates of root disease.

(c) For some link function h,

$$h[\mathrm{E}\{Y(\boldsymbol{s}) \mid b(\boldsymbol{s})\}] = \sum_{j=1}^{p} x_j(\boldsymbol{s})\beta_j + b(\boldsymbol{s}).$$

In practice, data are only available at the sampling sites $\boldsymbol{s}_i$, $i = 1, 2, \ldots, n$. Let Y_i, x_{i1}, x_{i2}, $\ldots$, x_{ip} be observations of the $(p+1)$ variables at site $\boldsymbol{s}_i$. Write $b_i = b(\boldsymbol{s}_i)$, $\boldsymbol{b} = (b_1, \ldots, b_n)$, and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. In many applications, predicting the random effects at unsampled sites is of great practical interest. For example, in the root disease example, a map of the random effects $b(\boldsymbol{s})$ across the field shows the severity of the disease and is helpful for efficiently treating the disease. Prediction of the random effects requires modeling the spatial dependence continuously. In recent years, there have been many works on modeling the spatial dependence continuously, including, among others, Handcock and Stein (1993), De Oliveira, Kadeem, and Short (1997), Ecker and Gelfand (1997), Diggle et al. (1998), Lahiri et al. (1999), Stein (1999), Sanso and Guenni (2000), and Wikle et al. (2001). Most of these works either do not consider interpolation of spatial non-Gaussian variables or do so by first transforming the variables to normality. An exception is that of Diggle et al. (1998).

We will focus on interpolation of random effects over a continuous spatial area when the observations are non-Gaussian. It is well known that the MMSE prediction for the random effect $b = b(\boldsymbol{s})$ at a site $\boldsymbol{s}$ is the conditional expectation $\mathrm{E}(b \mid \boldsymbol{Y})$. The MMSE prediction is particularly appropriate for spatial GLMM due to the following linear property analogous to linear kriging:

$$\mathrm{E}(b \mid \boldsymbol{Y}) = \sum_{i=1}^{n} c_i \mathrm{E}(b_i \mid \boldsymbol{Y}), \tag{1}$$

where $\mathrm{E}(b_i \mid \boldsymbol{Y})$ is the MMSE estimation of the realized random effect b_i and the coefficients c_i are such that $\Sigma_{i=1}^{n} c_i b_i$ equals $\mathrm{E}(b \mid \boldsymbol{b})$, the MMSE prediction of b given $\boldsymbol{b}$. In other words, these coefficients are the same as those in the MMSE prediction of $b(\boldsymbol{s})$ given $\boldsymbol{b}$. Hence, once the MMSE estimates of random effects are obtained at the sampling sites, the MMSE prediction for the random effect at any unsampled sites can be carried out as if the random effects were observable at the sampling sites.

It seems that equation (1) has not been used for prediction in spatial GLMMs. The objective of this current work is threefold: First, we establish (1) for the spatial GLMM. Second, given that parameter estimates are obtained by some method (we will review some of the methods in Section 2), we develop a Markov chain Monte Carlo (MCMC) method for estimating $\mathrm{E}(b_i \mid \boldsymbol{Y})$ that is implemented through the Metropolis–Hastings algorithm. Third, we develop a Monte Carlo version of the EM gradient algorithm, MCEMG for short. One advantage of the MCEMG is that it provides maximum likelihood estimates of parameters as well as the MMSE estimates of the realized random effects at the sampling sites. Hence, the MMSE prediction of the random effect $b(\boldsymbol{s})$ at any unsampled site $\boldsymbol{s}$ can be readily carried out linearly in light of (1). Although Monte Carlo versions of the EM algorithm or its variants have been used in the general GLMM context, almost all such works focus on estimation of parameters. However, for spatial GLMMs, estimation and prediction of random effects are usually an important goal, and the correlation structure introduced by the spatial random effects is more complex. It is these differences that warrant the investigation of applicability and performance of the Monte Carlo EM algorithm or its variants.

In Section 2, we describe the methodology and the EM and MCEMG algorithms. We establish (1) and show how to implement it through the Metropolis–Hastings algorithm. We present a simulation study in Section 3 and apply the method to the *Rhizoctonia* root rot data in Section 4. Some remarks and discussion are presented in the last section.

2. Methodology

2.1 *MMSE Prediction and the Metropolis–Hastings Algorithm*

In this subsection, we study prediction of random effects under an assumption that the parameters, β and θ, are known. The approach is based on the following general theorem.

THEOREM 1: *Let $b(\boldsymbol{s}), \boldsymbol{s} \in R^2$ be Gaussian with $\mathrm{E}\{b(\boldsymbol{s})\} = 0$ for all $\boldsymbol{s}$. If conditionally on $\{b(\boldsymbol{s}), \boldsymbol{s} \in R^2\}$, $\{Y(\boldsymbol{s}), \boldsymbol{s} \in R^2\}$ is an independent process and for each $\boldsymbol{s}$ the distribution of $Y(\boldsymbol{s})$ depends on $b(\boldsymbol{s})$ only, then for any $\boldsymbol{s}, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$,*

$$\mathrm{E}\{b(\boldsymbol{s}) \mid \boldsymbol{Y}\} = \sum_{i=1}^{n} c_i \mathrm{E}\{b(\boldsymbol{s}_i) \mid \boldsymbol{Y}\},$$

where the coefficients c_i are such that $\mathrm{E}\{b(\boldsymbol{s}) \mid b(\boldsymbol{s}_i),\ i = 1, \ldots, n\} = \Sigma_{i=1}^{n} c_i b(\boldsymbol{s}_i)$ and $\boldsymbol{Y} = (Y(\boldsymbol{s}_i),\ i = 1, \ldots, n)$.

Note that this theorem holds under conditions that are more general than the GLMM. Proof of the theorem is provided in the Appendix.

For any sampling site $\boldsymbol{s}_i$, $\mathrm{E}\{b(\boldsymbol{s}_i) \mid \boldsymbol{Y}\}$ cannot be given in closed form when $\boldsymbol{Y}$ is not Gaussian but can be approximated by Monte Carlo samples. We will generate Monte Carlo samples $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \ldots, \boldsymbol{b}^{(N)}$ from the conditional distribution $f_{\boldsymbol{b}|\boldsymbol{Y}}(\cdot \mid \boldsymbol{Y})$ through the Metropolis–Hastings (MH) algorithm so that, for any continuous function g, $\lim_{N\to\infty}(1/N)\Sigma_{m=1}^{N} g(\boldsymbol{b}^{(m)}) = \mathrm{E}\{g(\boldsymbol{b}) \mid \boldsymbol{Y}\}$. In particular, $\mathrm{E}\{b_i \mid \boldsymbol{Y}\} = \lim_{N\to\infty}(1/N)\Sigma_{m=1}^{N} b_i^{(m)}$, where $b_i^{(m)}$ is the ith element of $\boldsymbol{b}^{(m)}$. The MH algorithm is chosen due to its simplicity of implementation.

We refer to Chapter 1 of Gilks, Richardson, and Spoegelhalter (1996) for an introduction to the Metropolis–Hastings algorithm. If the candidate distribution is $f_{\boldsymbol{b}}(\cdot \mid \theta)$, the probability of accepting a new value $\boldsymbol{b}^*$ with the current value being $\boldsymbol{b}$ is the minimum of $\{f(\boldsymbol{b}^* \mid \boldsymbol{Y}, \beta, \theta) f_{\boldsymbol{b}}(\boldsymbol{b} \mid \theta)\}/\{f(\boldsymbol{b} \mid \boldsymbol{Y}, \beta, \theta) f_{\boldsymbol{b}}(\boldsymbol{b}^* \mid \theta)\}$ and one, and the ratio simplifies to $\Pi_{i=1}^n f(y_i \mid b_i^*, \beta)/f(y_i \mid b_i, \beta)$ due to the conditional independence. If we use the single-component Metropolis–Hastings algorithm, i.e., at each iteration, we only update a single component, say the kth component b_k, then the acceptance probability is further simplified to $\min\{1, f(y_k \mid b_k^*, \beta)/f(y_k \mid b_k, \beta)\}$. Note that $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$ has dependent Gaussian components; hence, to generate a new value for the kth component b_k while keeping other components unchanged, we need to sample from the conditional distribution of b_k given $b_j, j \neq k$, which is $\mathrm{N}(-\Sigma_{j \neq k} Q_{kj} b_j Q_{kk}^{-1}, Q_{kk}^{-1})$ when $\boldsymbol{b}$ is $\mathrm{MVN}(0, V)$, where Q_{kj} is the (k, j) element of the inverse of V.

For given parameters β and θ, we use the following single-component Metropolis–Hastings algorithm to generate Monte Carlo samples from $f_{\boldsymbol{b}|\boldsymbol{Y}}(\boldsymbol{b} \mid \boldsymbol{Y}, \beta, \theta)$:

```
Start with b^(0) = (0, ..., 0); set m=0.
Repeat{
for (k = 1 : n){
Generate a random value b_k^* from
N(-Σ_{j≠k} (Q_kj/Q_kk) b_j^(m), 1/Q_kk);
Generate a uniform(0, 1) random value U;
If U < min{1, f(y_k | b_k^*, β)/f(y_k | b_k, β)}, set b^(m) =
(b_1, ..., b_{k-1}, b_k^*, b_{k+1}, ..., b_n).
Otherwise, b^(m) stays unchanged.
}
Set b^(m+1) to be the current value of b^(m).
}
```

Note that here we take a sample only after each coordinate has been visited and the first N_0 burn-in samples should be discarded. Geyer (1992) suggested using an N_0 that is between 1% and 2% of the run length.

2.2 *Maximum Likelihood Estimation and the MCEMG Algorithm*

We consider maximum likelihood estimation of model parameters in this subsection. As in many other works, we assume the covariance function of the Gaussian process has a parametric form depending on some parameters θ of finite dimension, such as the exponential isotropic covariogram. Then, under the model described in the Introduction, the observed-data likelihood function is

$$L(\beta, \theta; \boldsymbol{Y}) = \int_{\mathbb{R}^n} \left\{ \prod_{i=1}^n f_{Y_i|\boldsymbol{b}}(y_i \mid \boldsymbol{b}, \beta) \right\} f_{\boldsymbol{b}}(\boldsymbol{b} \mid \theta) d\boldsymbol{b}.$$

The integral has a high dimension, and consequently it is intractable to find the MLE by directly maximizing L or $\ln L$. Several approaches have been proposed for the maximum likelihood estimation in GLMMs. Some are approximate inferences, as in Breslow and Clayton (1993) and Schall (1991), whose approaches are essentially equivalent to maximizing the joint distribution of $(\boldsymbol{Y}, \boldsymbol{b})$ with respect to the parameters and the random effects $\boldsymbol{b}$. Some incorporate Monte Carlo methods into the EM algorithm to obtain the maximum likelihood estimation (cf., Wei and Tanner (1990) and McCulloch (1994, 1997) for GLMM with independent random effects, Chan and Ledolter (1995) for time series models with latent correlated random effects, and Chan and Kuk (1997) for probit-linear mixed models with correlated random effects). In all these work, emphasis was given to inferences of parameters and not to estimation or prediction of random effects, and the random effects were not spatial.

The EM-type algorithm has become a standard procedure for estimation in GLMMs since the work of Dempster, Laird, and Rubin (1977). In an EM algorithm, the spatial random effects are considered missing data. The complete-data log-likelihood function is $\ln L_c(\beta, \theta; \boldsymbol{Y}, \boldsymbol{b}) = \ln f_{\boldsymbol{Y}|\boldsymbol{b}}(\boldsymbol{Y} \mid \boldsymbol{b}, \beta) + \ln f_{\boldsymbol{b}}(\boldsymbol{b} \mid \theta)$. The EM algorithm proceeds iteratively by maximizing the conditional expectation of the complete-data log likelihood $\mathrm{E}(\ln L_c \mid \boldsymbol{Y})$ at each iteration (the M-step), where the expectation is taken under the current value. We refer to McLachlan and Krishnan (1997) for an introduction to the EM algorithm and its variants. Some algorithms have been developed to speed up convergence of EM, one of which is the EM gradient (EMG) algorithm that substitutes a one-step Newton–Raphson algorithm for the M-step (Lange, 1995a,b). Write $\alpha = (\beta, \theta)$ and

$$\mathcal{I}(\alpha) = -\mathrm{E}\left(\frac{\partial^2 \ln L_c}{\partial \alpha \partial \alpha'} \mid \boldsymbol{Y} \right),$$
$$S(\alpha) = \mathrm{E}\left(\frac{\partial \ln L_c}{\partial \alpha} \mid \boldsymbol{Y} \right), \tag{2}$$

where the derivatives and expectations are evaluated at α. The EMG algorithm updates the estimates by $\alpha^{(m+1)} = \alpha^{(m)} + \mathcal{I}(\alpha^{(m)})^{-1} S(\alpha^{(m)})$, which breaks down to two since the information matrix $\mathcal{I}$ is clearly block diagonal,

$$\beta^{(m+1)} = \beta^{(m)} - \left[\mathrm{E}\left\{ \frac{\partial^2 \ln f\left(\boldsymbol{Y} \mid \boldsymbol{b}, \beta^{(m)}\right)}{\partial \beta \partial \beta'} \mid \boldsymbol{Y} \right\} \right]^{-1} \times \mathrm{E}\left\{ \frac{\partial \ln f_{\boldsymbol{Y}|\boldsymbol{b}}\left(\boldsymbol{Y} \mid \boldsymbol{b}, \beta^{(m)}\right)}{\partial \beta} \mid \boldsymbol{Y} \right\}, \tag{3}$$

$$\theta^{(m+1)} = \theta^{(m)} - \left[\mathrm{E}\left\{ \frac{\partial^2 \ln f\left(\boldsymbol{b} \mid \theta^{(m)}\right)}{\partial \theta \partial \theta'} \mid \boldsymbol{Y} \right\} \right]^{-1} \times \mathrm{E}\left\{ \frac{\partial \ln f_{\boldsymbol{b}}\left(\boldsymbol{b} \mid \theta^{(m)}\right)}{\partial \theta} \mid \boldsymbol{Y} \right\}, \tag{4}$$

where all conditional expectations are evaluated under the current parameter values $\beta^{(m)}$ and $\theta^{(m)}$. This iterative procedure continues until convergence is achieved. The parameter θ usually has nonnegative elements. We can meet these parameter constraints by halving the step size, a technique commonly employed in practice (e.g., Zimmerman and Zimmerman, 1991; Breslow and Clayton, 1993).

We note that, if the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{b}$ is from the exponential family and the link is canonical, the derivatives in (3) can be given in closed form (McCullagh and Nelder, 1989, p. 40–42). In particular, if $\boldsymbol{Y} \mid \boldsymbol{b}$ is binomial or

Poisson with a canonical link function,

$$\frac{\partial \ln f(\boldsymbol{Y} \mid \boldsymbol{b}, \beta)}{\partial \beta} = X'(\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{b})),$$
$$\frac{\partial^2 \ln f(\boldsymbol{Y} \mid \boldsymbol{b}, \beta)}{\partial \beta \partial \beta'} = -X'V(\boldsymbol{Y} \mid \boldsymbol{b})X, \qquad (5)$$

where $V(\boldsymbol{Y} \mid \boldsymbol{b})$ is the diagonal matrix of the conditional variance matrix of $\boldsymbol{Y}$ given $\boldsymbol{b}$ and $X = (x_{ij})$ is the design matrix. The derivatives in (4) can also be given in closed form since $\boldsymbol{b}$ is multivariate normal (see Mardia and Marshall (1984) or relevant results of matrix theory, e.g., Graybill (1983, Chapter 10)):

$$\frac{\partial \ln f(\boldsymbol{b} \mid \theta)}{\partial \theta_i} = -\frac{1}{2}\mathrm{tr}\left(V^{-1}V_i\right) + \frac{1}{2}\boldsymbol{b}'\left(V^{-1}V_iV^{-1}\right)\boldsymbol{b},$$
$$\frac{\partial^2 \ln f(\boldsymbol{b} \mid \theta)}{\partial \theta_i \partial \theta_j} = -\frac{1}{2}\mathrm{tr}\left(V^{-1}V_{ij} - V^{-1}V_iV^{-1}V_j\right) - \frac{1}{2}\boldsymbol{b}'V^{ij}\boldsymbol{b},$$

where V^{-1} is the inverse of the variance matrix $V(\theta)$ of $\boldsymbol{b}$ and

$$V_i = \frac{\partial V}{\partial \theta_i},$$
$$V_{ij} = \frac{\partial^2 V}{\partial \theta_i \partial \theta_j},$$
$$V^{ij} = \frac{\partial^2 V^{-1}}{\partial \theta_i \partial \theta_j} = V^{-1}\left(V_iV^{-1}V_j + V_jV^{-1}V_i - V_{ij}\right)V^{-1}.$$

The conditional expectations in (3) and (4) cannot be calculated in closed form but can be approximated using the Monte Carlo samples $\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(N)}$ from the Metropolis–Hastings algorithm under the current estimates $\beta^{(m)}$ and $\theta^{(m)}$. For example,

$$-\mathrm{E}\left\{\frac{\partial^2 \ln f(\boldsymbol{Y} \mid \boldsymbol{b}, \beta^{(m)})}{\partial \beta \partial \beta'} \mid \boldsymbol{Y}\right\} = X'\mathrm{E}\{V(\boldsymbol{Y} \mid \boldsymbol{b}) \mid \boldsymbol{Y}\}X$$
$$\approx 1/N \sum_{m=1}^{N} X'V\left(\boldsymbol{Y} \mid \boldsymbol{b}^{(m)}\right)X.$$

Incorporating this Monte Carlo technique into the EMG algorithm results in the MCEMG algorithm. Lange (1995b) noted that the local properties of the EM gradient algorithm are almost identical to those of the EM algorithm. The Monte Carlo version should inherit this property.

We can choose a starting value for β by first fitting a GLMM with i.i.d. random effects. From the resulting estimates of the random effects, we calculate the empirical variogram

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{N(h)}\left(\hat{b}(\boldsymbol{s}_i) - \hat{b}(\boldsymbol{s}_j)\right)^2, \qquad h > 0,$$

where $N(h) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : |\boldsymbol{s}_i - \boldsymbol{s}_j| = h\}$ and $|N(h)|$ is the number of distinct pairs in $N(h)$. We then plot this empirical variogram, which may help us gain some insight into the parametric form of the variogram and choose initial values of the variogram parameters. Care must be given to the choice of parameters of the variogram since the variability of estimated random effects is smaller than the variability of the unobservable random effects. Therefore, the empirical variogram from $\hat{\boldsymbol{b}}$ has a smaller sill than that of $\boldsymbol{b}$. We will further discuss the choice of initial values in Section 3.

2.3 *Information Matrix*

The observed information matrix is defined as the negative of the second derivative of the observed-data log likelihood with respect to the parameter $\alpha = (\beta, \theta)$, i.e., $I_Y(\beta, \theta) = -\partial^2 \ln L/\partial\alpha\partial\alpha'$, where L is the observed-data likelihood. It is easier to compute than the Fisher information matrix $\mathrm{E}(I_Y(\beta, \theta))$ and in most cases is a more appropriate measure of information (Efron and Hinkley, 1978). Louis (1982) obtained the following result in the EM framework:

$$I_Y(\beta, \theta) = \mathcal{I}(\beta, \theta) - \mathrm{E}\{S_c(\beta, \theta; \boldsymbol{Y}, \boldsymbol{b})S_c'(\beta, \theta; \boldsymbol{Y}, \boldsymbol{b}) \mid \boldsymbol{Y}\}$$
$$+ \mathrm{E}\{S_c(\beta, \theta; \boldsymbol{Y}, \boldsymbol{b}) \mid \boldsymbol{Y}\}\mathrm{E}\{S_c'(\beta, \theta; \boldsymbol{Y}, \boldsymbol{b}) \mid \boldsymbol{Y}\}, \qquad (6)$$

where $\mathcal{I}(\beta, \theta)$ is defined in (2), $S_c(\beta, \theta; \boldsymbol{Y}, \boldsymbol{b})$ is the first derivative of the complete-data log likelihood, and the expectations are all taken under the parameters (β, θ). The observed information matrix needs only to be calculated at the last step in the EM or MCEMG algorithm when it converges, for which the last term in (6) is zero. Again, the observed information is obtainable via the MCMC technique.

3. A Simulation Study

In this section, we present a simulation study to discuss the choice of initial values, determination of the Monte Carlo sample size, and some diagnostic techniques. We simulated data from the following model on a 15×15 lattice: $Y_{ij} \mid \boldsymbol{b}$ is binomial with $n_{ij} = 10$ and $p_{ij} = 1 - 1/\exp(-2 + 0.15i + b_{ij})$, $i, j = 1, \ldots, 15$, and $\boldsymbol{b} = (b_{ij})$ is from a stationary Gaussian process with an isotropic exponential variogram, $\gamma(h) = 0.5 + 2(1 - \exp(-h/5))$ for $h > 0$ and $\gamma(0) = 0$. The S-Plus function `rfsim` was used to simulate the Gaussian b_{ij}. Conditional on b_{ij}, the binomial random variables Y_{ij} were simulated using the S-Plus function `rbinom`. Write $\beta = (-2, 0.15)$ and $\theta = (0.5, 2, 5)$.

We first fitted the data by a binomial mixed model with i.i.d. random effects. Applying the MCEMG algorithm, we obtained the estimate $\tilde{\beta} = (-1.6843, 0.1386)$ and the estimates of the realized random effects. The empirical variogram from the estimates of the random effects was calculated, from which we obtained an estimate for θ, $\tilde{\theta} = (0.6209, 0.5856, 2.4433)$, through the least squares method (Cressie, 1993, p. 94). We then used $\tilde{\beta}$ and $\tilde{\theta}$ as initial values to run MCEMG for the spatial GLMM. The Monte Carlo sample size was 2000 and the burn-in length was 200, i.e., the first 200 data samples were ignored and the last 2000 retained. The MCEMG estimates were then $\hat{\beta} = (-1.9161, 0.1626)$ and $\hat{\theta} = (0.6029, 1.2187, 2.6289)$, and their estimated standard deviations were obtained as the square roots of the diagonal elements of the inverse observed information matrix, (0.1004, 0.0101, 0.1530, 0.3431, 1.4412). Estimators for θ have larger standard deviations, as in general linear models with spatially correlated error terms.

Because $\hat{\theta}_3$, the range, is usually much larger than the other parameters, we did not use a criterion on the absolute difference of estimates in two consecutive iterations. Instead, we declared convergence if the absolute difference between $\mathrm{E}\{\ln L_c(\alpha^{(m+1)}; \boldsymbol{Y}, \boldsymbol{b}) \mid \boldsymbol{Y}\} = \max_\alpha \mathrm{E}\{\ln L_c(\alpha; \boldsymbol{Y}, \boldsymbol{b}) \mid \boldsymbol{Y}\}$ and $\mathrm{E}\{\ln L_c(\alpha^{(m)}; \boldsymbol{Y}, \boldsymbol{b}) \mid \boldsymbol{Y}\}$ was less than some specified value, δ, where the conditional expectations were under $\alpha^{(m)} = (\beta^{(m)}, \theta^{(m)})$. This criterion is based on the following

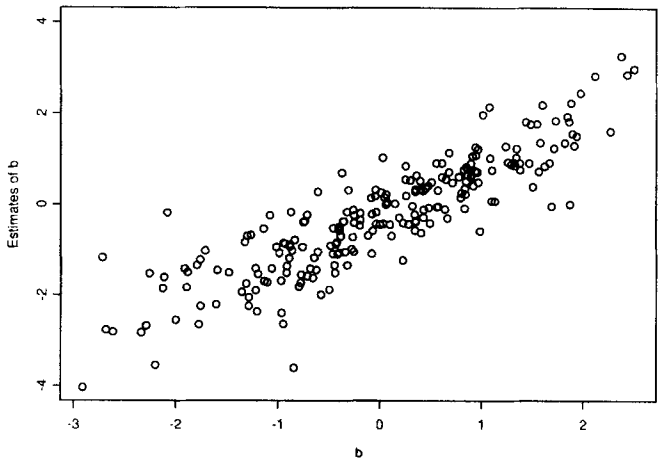

Figure 2. Scatter plots of the estimated (vertical axis) and realized values of random effects (horizontal axis).

property of EM: $\alpha^{(m)}$ is an MLE if and only if $L(\alpha^{(m+1)}) = L(\alpha^{(m)})$ and the equality holds if and only if $\mathrm{E}\{\ln L_c(\alpha^{(m+1)}; \boldsymbol{Y}, \boldsymbol{b}) \mid \boldsymbol{Y}\} = \mathrm{E}\{\ln L_c(\alpha^{(m)}; \boldsymbol{Y}, \boldsymbol{b}) \mid \boldsymbol{Y}\}$ (Robert and Casella, 1999, Theorem 5.3.4, p. 214). Clearly, if L is unimodal, this criterion is equivalent to the one on the absolute difference of estimates.

The estimated random effects ($\hat{\boldsymbol{b}}$) were compared with the realized random effects ($\boldsymbol{b}$) in two-ways: The scatter plot of $\hat{\boldsymbol{b}}$ and $\boldsymbol{b}$ presented in Figure 2 shows how close $\hat{\boldsymbol{b}}$ is to $\boldsymbol{b}$, and the gray-level images of both $\hat{\boldsymbol{b}}$ and $\boldsymbol{b}$ presented in Figure 3 reveal how well $\hat{\boldsymbol{b}}$ preserves the overall pattern of the realized random effects $\boldsymbol{b}$. Overall, $\hat{\boldsymbol{b}}$ resembles $\boldsymbol{b}$ well.

Also plotted in Figure 3 are the two fitted variograms, one corresponding to the MCEMG estimate $\hat{\theta}$ and the other to the

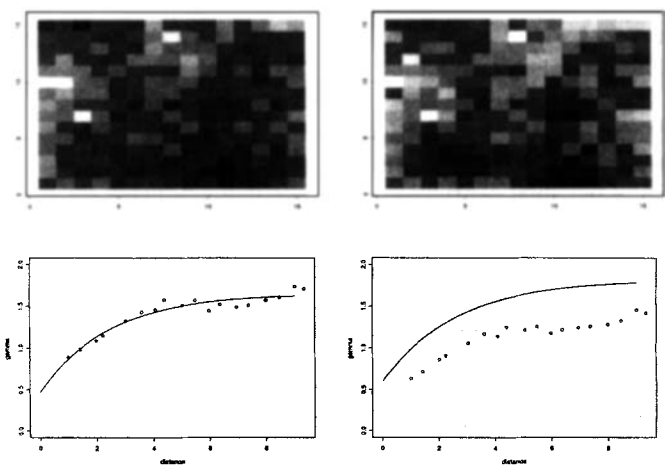

Figure 3. Gray-level plots (upper) and variograms(lower) of the realized random effects (left) and estimated random effects (right). Circles are the values of the empirical variograms. The solid line on the lower left is the least squares fit of the empirical variogram of the realized random effects. The solid line on the lower right is the estimated variogram corresponding to estimates from the MCEMG algorithm.

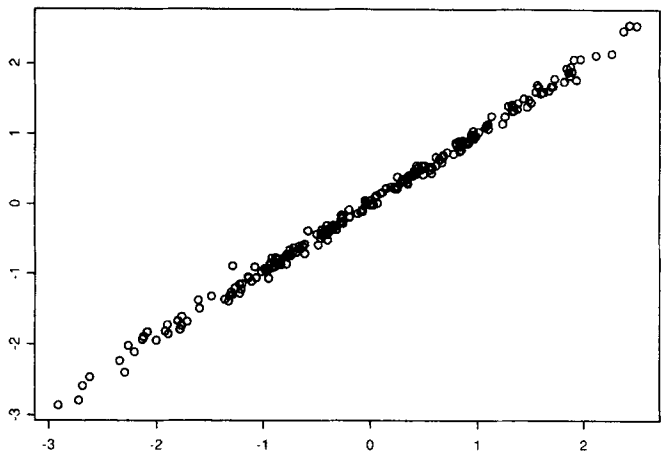

Figure 4. Comparison of estimated random effects corresponding to two sets of θ: the horizontal axis corresponds to $\hat{\theta} = (0.6029, 1.2187, 2.6289)$ and the vertical axis to $(0.7626, 6.6815, 26.2376)$; β is fixed at $(-1.9161, 0.1623)$.

least-squares estimates (0.4663, 1.1852, 2.4154), which were obtained from the realized random effects $\boldsymbol{b}$. When judging performance of MCEMG, we need to bear in mind that a particular set of realized values of random effects was used in MCEMG. We shall not demand that the estimate of θ given by MCEMG outperforms the estimate of θ directly from the realized random effects $\boldsymbol{b}$. Considering this, we believe the MCEMG estimates are satisfactory. The empirical variograms corresponding to the realized random effects and the estimated random effects are also plotted in Figure 3, from which we see the latter has a smaller sill (the limiting upper bound of the variogram). This is due to the fact that $\mathrm{E}(b \mid \boldsymbol{Y})$ has a smaller variance than b.

We find it interesting that the estimates of $\boldsymbol{b}$ are not greatly affected by $\hat{\theta}$, especially when $\hat{\theta}_1$ and $\hat{\theta}_2$ are larger than the true values. For example, we used the true value of θ as an initial value while keeping the same initial value for β. It did not yield converging estimates for θ (indeed, $\hat{\theta}$ increased in each iteration) but produced estimates for β that were always close to the true values after a few iterations. At the 30th iteration, the estimates were $\hat{\beta} = (-1.9750, 0.1675)$ and $\hat{\theta} = (0.7626, 6.6816, 26.2376)$. We compared in Figure 4 the estimated random effects corresponding to these estimates with those obtained previously that corresponded to the estimates by MCEMG. We see that the two sets of estimates of $\boldsymbol{b}$ are very close. This might suggest that we start with a large sill and run a few iterations of the MCEMG algorithm and obtain the estimate of random effects. We then obtain the empirical variogram of these random effects estimates, which should not only help check the validity of the parametric form of the variogram but also provide estimates of variogram parameters. These estimates will then be used as initial values to run the MCEMG algorithm. We certainly can explore this approach if the initial values obtained from the GLMM with i.i.d. random effects fail to result in convergence.

To monitor convergence in the EM or MCEMG, several authors have suggested plotting estimates at each iteration (e.g., Wei and Tanner, 1990; Chan and Ledolter, 1995). Con-

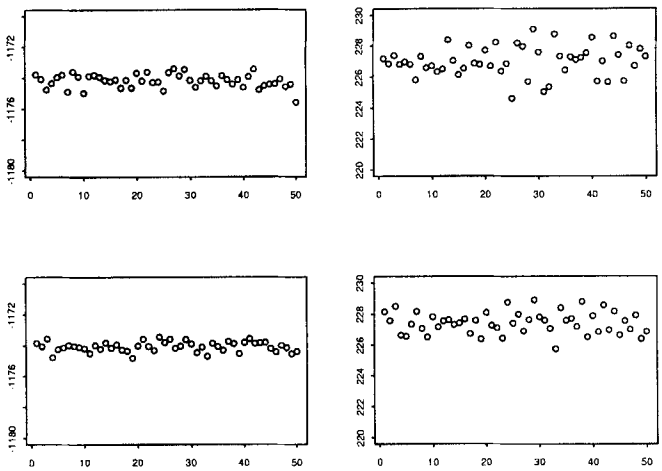

Figure 5. Fifty estimates of $\mathrm{E}(\ln f(\boldsymbol{Y} \mid \boldsymbol{b}) \mid \boldsymbol{Y})$, each based on a Monte Carlo sample size $N = 2000$ (upper left) and $N = 5000$ (lower left); 50 estimates of $\mathrm{E}(\ln f(\boldsymbol{b} \mid \theta) \mid \boldsymbol{Y})$ each based on $N = 2000$ (upper right) and $N = 5000$ (lower right). The conditional expectations correspond to $\hat{\beta} = (-1.9161, 0.1623)$ and $\hat{\theta} = (0.6029, 1.2187, 2.6289)$.

vergence is then indicated by small random fluctuations of estimates from iteration to iteration. In our algorithm, the variations of estimates from one iteration to another come from two sources: One is the Monte Carlo approximation and the another is the iterative nature of the EM algorithm. To specifically see whether the Monte Carlo sample size in the Metropolis–Hastings algorithm is adequately large, we can compute and plot the observed log-likelihood function $\mathrm{E}(L_c \mid \boldsymbol{Y})$ after convergence is achieved in the MCEMG algorithm, where the Monte Carlo estimate for $\mathrm{E}(L_c \mid \boldsymbol{Y})$ is based on the converged estimates $\hat{\beta}$ and $\hat{\theta}$. A large fluctuation would indicate the Monte Carlo sample size is not large enough to yield numerically stable estimates. For the simulated data, we generated 50 sets of Monte Carlo samples, each of size 2000, and plotted the estimates $\hat{\mathrm{E}}(L_c \mid \boldsymbol{Y})$ in Figure 5. $\hat{\mathrm{E}}(\ln f(\boldsymbol{Y} \mid \boldsymbol{b}) \mid \boldsymbol{Y})$ and $\hat{\mathrm{E}}(\ln f(\boldsymbol{b}) \mid \boldsymbol{Y})$ lie in $(-1175.58, -1173.393)$ and $(224.82, 229.09)$, respectively. When the Monte Carlo sample sizes are increased to 5000, the two ranges become $(-1174.80, -1173.44)$ and $(226.03, 228.87)$. It seems that a Monte Carlo size between 2000 and 5000 is sufficient.

Since the Monte Carlo samples are correlated, the standard errors of means of the Monte Carlo samples are more directly measured by the effective sample sizes than by the chain length. For each component $b_i = b(\boldsymbol{s}_i)$, the corresponding effective sample size is the chain length divided by the autocorrelation time, τ, which is defined to be $1 + 2\Sigma_{k=1}^{\infty} \rho(k)$, where $\rho(k)$ is the autocorrelation of MCMC sample $\{b_i^{(m)}, m = 1, \ldots, N\}$ at lag k (cf., Hastings, 1970; Sargent, Hodges, and Carlin, 2000). To estimate the autocorrelation time, the sum was cut off at a k where $\rho(j), j > k$ seemed to fall between -0.075 and 0.075. The autocorrelation times for the 225 components of $\boldsymbol{b}$ ranged from 1.209 to 4.945 with a mean 2.361 and the effective sample sizes from 404.4 to 1654 with a mean 949.5 when the chain length was 2000. The acceptance rate of the Metropolis–Hastings algorithm was persistently between 45 and 58%.

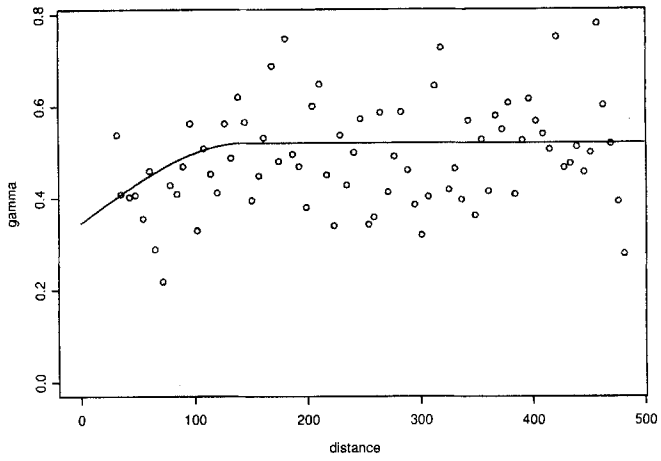

Figure 6. Empirical variogram of estimated random effects and fitted variogram by the MCEMG algorithm for the root rot data.

We repeated the simulation study 30 times in order to estimate the biases of the estimators. We increased the MC sample size to 5000 to reduce variations of estimates due to MC sampling. The biases of $(\hat{\beta}, \hat{\theta})$ are (0.035, −0.026, 0.063, −0.231, −1.02) and the standard deviations of the estimates are (0.176, 0.040, 0.066, 0.337, 1.067). It seemed that biases of $\hat{\beta}$ were negligible. In most cases, convergence was achieved in less than 20 iterations. Each iteration took about 105 seconds on a 733-MHz Pentium III with a 128-MB SDRAM at 133 MHz, and most of the computing time in each iteration was on updating $\hat{\theta}$. In some cases, the algorithm failed to converge because $\hat{\theta}_2$ or $\hat{\theta}_3$ was either always increasing or decreasing without converging. The constant δ was fixed at one in all the simulations. The algorithm was run in S-Plus and called many Fortran subroutines for loops such as for the Metropolis–Hastings sampling.

4. Analysis on the Incidence Rates of *Rhizoctonia* Root Rot

We apply the spatial GLMM with the number of infected crown roots as the binomial response variable and assume it has a binomial distribution, with the index n_i being the total number of crown roots at site $\boldsymbol{s}_i$ and the parameter p_i being $\exp(\beta + b_i)/(1 + \exp(\beta + b_i))$, where b_i's are assumed to be Gaussian isotropically stationary with a spherical variogram with parameter $\theta = (\theta_1, \theta_2, \theta_3)$, where $\theta_i, i = 1, 2, 3$, are the nugget, partial sill, and range, respectively. We chose a spherical variogram because the empirical variogram of barley yields, which is not included in this article, seemed to be flat after a distance, and it was believed that yield and the root rot were significantly correlated. Using the MCEMG algorithm with a Monte Carlo sample size 5000, we obtained the estimates $\hat{\beta} = -1.6152$ and $\hat{\theta} = (0.3451, 0.1754, 145.11)$, with estimated standard deviations $(0.0023, 0.0898, 0.1086, 73.33)$. Figure 6 shows the fitted variogram corresponding to $\hat{\theta}$ and the empirical variogram obtained from the estimated random effects and seems to suggest that the spherical variogram is a reasonable choice for the data. Using (1) and the estimates of b_i and θ, we obtain predictions for random effects at 3111

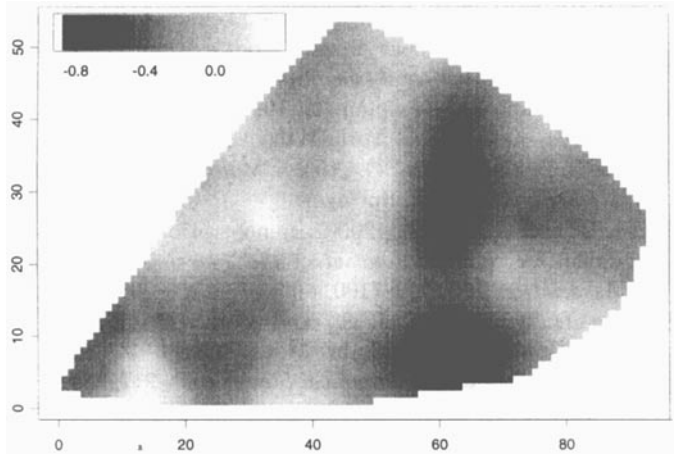

Figure 7. Map of interpolated random effects at 3111 sites for the root rot data.

sites, which are mapped in Figure 7. The high-incident zones correspond to low-yield zones.

We also generated 50 sets of Monte Carlo samples, each of which corresponded to the estimates $\hat{\beta}$ and $\hat{\theta}$ and having size 5000. With each of the 50 sets, we calculated $\hat{\mathrm{E}}(\ln f(\boldsymbol{Y} \mid \boldsymbol{b}) \mid \boldsymbol{Y})$ and $\hat{\mathrm{E}}(f(\boldsymbol{b} \mid \boldsymbol{Y}))$, which ranged from -1175.57 to -1173.39 and 116.24 to 119.15. A Monte Carlo sample size of 5000 seemed adequate.

5. Discussion and Conclusion

We have developed the MMSE prediction of random effects in a GLMM, which can be implemented through the Metropolis–Hastings algorithm. Once parameter estimates are obtained from some method, not necessarily the MCEMG, prediction can be done linearly in light of (1). However, the MCEMG algorithm provides estimates of parameters as well as MMSE estimates of the random effects on sampling sites. Simulation results show MCEMG works reasonably well for spatial GLMM.

Determining the parametric form of the variogram in a spatial GLMM is a difficult problem and deserves further study. When a spatial variable is observable, a frequently used approach in geostatistics is to calculate and plot the empirical variogram and then choose a parametric variogram to be fitted via the least-squares or maximum-likelihood techniques. Despite its popularity, this approach also faces criticism (Stein, 1999). After all, it cannot be directly applied to the spatial GLMM since the random effects are not observable. Diggle et al. (1998) tried to approximate the functional relationship between the variogram of the response variable in a spatial GLMM and that of the unobservable random effects. This approximation will become more complex when sample sizes are unequal. Stein (1999) favored fitting a Matérn variogram through maximum-likelihood techniques for Gaussian data. There is a lack of an adequate validation technique for fitting a variogram. Cross-validation may seem appealing but needs to be further studied in order to be appropriately used for confirmatory data analysis in spatial models (Cressie, 1993, p. 104). In this article, we plot the empirical covariogram calculated from the estimated random effects after convergence of MCEMG and compare it with the fitted covariogram by MCEMG. It is expected that the empirical variogram has a smaller sill than the fitted one. A severe discrepancy between the two variograms might suggest the parametric form assumed at the first place is not appropriate.

Restricted maximum likelihood (REML) estimation is sometimes preferred in general linear mixed models and particularly in spatial regression with normal errors (e.g., Zimmerman and Zimmerman, 1991; Cressie and Lahiri, 1996) to estimate the variance–covariance parameter θ since the MLE of θ is usually biased. In a general linear mixed model, REML estimation applies maximum likelihood estimation to error contrasts so that the distribution of the error contrasts depends only on θ. Breslow and Clayton (1993) used REML to estimate variance component parameters in a GLMM by introducing a working vector (also called the adjusted dependent variable; McCullagh and Nelder, 1989, p. 40; Schall, 1991) to linearize the response variable $\boldsymbol{Y}$ in the GLMM. This results in approximate inferences for the GLMM. It is not immediately clear how to accommodate REML in GLMMs without introducing a working vector, as in our approach in this article. Also not included in this work is the calculation of the mean-squared prediction error. It is possible to obtain mean-squared prediction errors in the spatial GLMM framework, but the length of this article prevents inclusion of it here. We will explore it in a separate article.

Résumé

Nous utilisons des modèles linéaires généralisés mixtes (GLMM) spatiaux pour modéliser des variables spatialisées non gaussiennes observées à des positions aléatoires dans une région continue. Dans de nombreuses applications, la prédiction des effets aléatoires au sein d'un GLMM spatial est d'un grand intérêt pratique. Nous démontrons que la prédiction de l'erreur quadratique moyenne minimale (MMSE) peut être faite linéairement dans des GLMM spatiaux, d'une façon analogue au krigeage linéaire. Nous développons une version Monte Carlo de l'algorithme du gradient EM pour l'estimation par maximum de vraisemblance des paramètres du modèle. Un sous-produit de cette approche est qu'elle fournit aussi des estimation de la MMSE pour les effets aléatoires réalisés aux points d'échantillonnage. La méthode est illustrée par une étude par simulation; elle est aussi appliquée à des données réelles de phytopathologie racinaire, ce qui permet d'obtenir une carte de la sévérité de la maladie facilitant des pratiques agronomiques de précision.

References

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

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

Chan, J. S. K. and Kuk, A. Y. C. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics* **53,** 86–97.

Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association* **90,** 242–252.

Cook, R. J. (1992). Wheat root health management and environmental concern. *Canadian Journal of Plant Pathology* **14,** 76–85.

Cook, R. J. and Haglund, W. A. (1991). Wheat yield depression association with conservation tillage caused by root pathogens in the soil not phytotoxins from the straw. *Soil Biology and Biochemistry* **23,** 125–1132.

Cressie, N. (1993). *Statistics for Spatial Data,* revised edition. New York: Wiley.

Cressie, N. and Lahiri, S. N. (1996). Asymptotics for REML estimation of spatial covariance parameters. *Journal of Statistical Planning and Inference* **50,** 327–341.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

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

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

Ecker, M. and Gelfand, A. (1997). Bayesian variogram modeling for an itropic spatial process. *Journal of Agricultural, Biological, and Environmental Statistics* **2,** 347–369.

Efron, B. and Hinkley, D. V. (1978). The observed versus expected information. *Biometrika* **65,** 457–487.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science* **7,** 473–511.

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

Graybill, F. A. (1983). *Matrices with Applications in Statistics,* 2nd edition. Belmont, California: Wadsworth.

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

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

Lahiri, S., Kaiser, M., Cressie, N., and Hsu, N. (1999). Prediction of spatial cumulative distribution functions using subsampling (with discussion). *Journal of the American Statistical Association* **94,** 86–110.

Lange, K. (1995a). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society, Series B* **57,** 425–437.

Lange, K. (1995b). A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica* **5,** 1–18.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44,** 226–233.

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

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

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association* **89,** 330–335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92,** 162–170.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* New York: Wiley.

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

Sanso, B. and Guenni, L. (2000). A nonstationary multisite model for rainfall. *Journal of the American Statistical Association* **95,** 1089–1100.

Sargent, D. J., Hodges, J. S., and Carlin, B. P. (2000). Structured Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* **9,** 217–234.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **40,** 719–727.

Shiryayev, A. N. (1984). *Probability.* New York: Springer-Verlag.

Stein, M. (1999). *Interpolation of Spatial Data.* New York: Springer.

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association* **92,** 607–617.

Wei, G.C.G. and Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85,** 699–704.

Wikle, C., Milliff, R., Nychka, D., and Berlinger, M. (2001). Spatial-temporal heirarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association* **96,** 382–397.

Zimmerman, D. L. and Zimmerman, M. B. (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics* **23,** 77–91.

Received April 2001. *Revised November* 2001. *Accepted November* 2001.

Appendix

Proof of Theorem 1

Let $\boldsymbol{s} \neq \boldsymbol{s}_i$ for all $i = 1, 2, \ldots, n$. Denote $b_0 = b(\boldsymbol{s})$ and the joint probability density function of $(b_0, \boldsymbol{b}, \boldsymbol{Y})$ by $f_{b_0,\boldsymbol{b},\boldsymbol{Y}}(b_0, \boldsymbol{b}, \boldsymbol{y})$, $b_0 \in \mathbb{R}$, $\boldsymbol{b}, \boldsymbol{y} \in \mathbb{R}^n$. Due to the model formation, the conditional distribution of $\boldsymbol{Y}$ given $\{b(\boldsymbol{s}), \boldsymbol{s} \in R^2\}$ is the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$. This implies $f_{\boldsymbol{Y}|b_0,\boldsymbol{b}}(\boldsymbol{y} \mid b_0, \boldsymbol{b}) = f_{\boldsymbol{Y}|\boldsymbol{b}}(\boldsymbol{y} \mid \boldsymbol{b})$. Therefore,

$$\begin{aligned} f_{b_0,\boldsymbol{b},\boldsymbol{Y}}(b_0, \boldsymbol{b}, \boldsymbol{y}) &= f_{\boldsymbol{Y}|b_0,\boldsymbol{b}}(\boldsymbol{y} \mid b_0, \boldsymbol{b}) f_{b_0,\boldsymbol{b}}(b_0, \boldsymbol{b}) \\ &= f_{\boldsymbol{Y}|\boldsymbol{b}}(\boldsymbol{y} \mid \boldsymbol{b}) f_{b_0,\boldsymbol{b}}(b_0, \boldsymbol{b}) \\ &= f_{\boldsymbol{b},\boldsymbol{Y}}(\boldsymbol{b}, \boldsymbol{y}) f_{b_0|\boldsymbol{b}}(b_0 \mid \boldsymbol{b}). \end{aligned}$$

Dividing both sides by $f_{\boldsymbol{b},\boldsymbol{Y}}(\boldsymbol{b}, \boldsymbol{y})$, we obtain $f_{b_0|\boldsymbol{b},\boldsymbol{Y}}(b_0 \mid \boldsymbol{b}, \boldsymbol{y}) = f_{b_0|\boldsymbol{b}}(b_0 \mid \boldsymbol{b})$ and consequently $\mathrm{E}(b_0 \mid \boldsymbol{b}, \boldsymbol{Y}) = \mathrm{E}(b_0 \mid \boldsymbol{b}) = \Sigma_{i=1}^{n} c_i b_i$ for some appropriate constants c_i. Equation (1) then follows from the well-known fact $\mathrm{E}(b_0 \mid \boldsymbol{Y}) = \mathrm{E}\{\mathrm{E}(b_0 \mid \boldsymbol{b}, \boldsymbol{Y}) \mid \boldsymbol{Y}\}$ (cf., Shiryayev, 1984, p. 214).