A Composite Likelihood Approach to Binary Spatial Data

Patrick J. Heagerty; Subhash R. Lele

Journal of the American Statistical Association, Vol. 93, No. 443 (Sep., 1998), 1099-1111.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199809%2993%3A443%3C1099%3AACLATB%3E2.0.CO%3B2-Y

Journal of the American Statistical Association is currently published by American Statistical Association.

A Composite Likelihood Approach to Binary Spatial Data

Patrick J. HEAGERTY and Subhash R. LELE

Conventional geostatistics addresses the problem of estimation and prediction for continuous observations. But in many practical applications in public health, environmental remediation, or ecological research the most commonly available data are in the form of counts (e.g., number of cases) or indicator variables denoting above or below threshold values. Also, in many situations it is less expensive to obtain an imprecise categorical observation than to obtain precise measurements of the variable of interest (such as a contaminant). This article proposes a computationally simple method for estimation and prediction using binary or indicator data in space. The proposed method is based on pairwise likelihood contributions, and the large-sample properties of the estimators are obtained in a straightforward manner. We illustrate the methodology through application to indicator data related to gypsy moth defoliation in Massachusetts.

KEY WORDS: Empirical Bayes; Estimating function; Hierarchical model; Indicator kriging; Iterated conditional modes; Latent variables.

1. INTRODUCTION

Classical statistical models for regression and prediction with spatial data are useful in situations where data are approximately Gaussian and can be modeled using a linear mean structure. Such models and the accompanying methods for inference are well studied and understood (for thorough reviews see Cressie 1991; Haining 1989). But application of these methods may be misleading in situations where the data are clearly non-Gaussian, such as with categorical or count data. Indicator kriging (Journel 1983) for binary data in space grew out of the need to extend the Gaussian methods. Disjunctive kriging (Matheron 1976) is another approach that uses nonlinear functions of the response for prediction. Trans-Gaussian kriging (Cressie 1991, pp. 137–138) accommodates non-Gaussian data by using marginal transforms of the response. What these approaches share is the goal of prediction through direct modeling of joint and conditional distributions of the response. But in the past decade, since the publication of Clayton and Kaldor (1987), there has been intense development of hierarchical models (Carlin and Louis 1996) for both longitudinal and spatial data. These models build the joint probability structure through specification of an observational level model that may depend on unobserved variables and a hidden or latent level model for the unobserved quantities. Inference procedures have been developed based on approximate maximum likelihood (Clayton and Kaldor 1987), penalized quasi-likelihood (Breslow and Clayton 1993), Markov chain Monte Carlo methods (Besag, York, and Mollie 1991; Diggle, Moyeed, and Tawn 1997; Waller, Carlin, Xio, and Gelfand 1997), and estimating functions (Yasui and Lele 1997).

Most of these methods can have practical difficulty handling large datasets due to computational demands. But many of the spatial datasets in public health, ecological pest control, or environmental remediation science contain thousands of observations. Also, obtaining categorical information such as an indicator that the amount of contamination is above or below a threshold, or that the abundance of a species is low, medium, or high can be substantially cheaper than obtaining precise response measurements. Therefore, it is often possible to obtain categorical information at many locations, and with a large dataset it is often computational feasibility rather than statistical efficiency that dominates analytical concerns. Our goal is not to supplant the careful collection of precise spatial data, but rather to recognize the availability of potentially informative spatial categorical datasets and to propose an alternative practical method for regression analysis and prediction that use these data. This article proposes a method of inference based on the concept of composite likelihood (Lindsay 1988) for the estimation of spatial hierarchical model parameters. Our method is computationally simple and can handle large amounts of data, and theoretical results such as consistency and asymptotic normality of the resultant estimators follow in a reasonable fashion. This article focuses on the analysis of binary data in spatial situations. Extensions of the composite likelihood approach to other categorical responses and to space–time data are feasible and a primary goal for future work.

In the next section we discuss a practical situation related to pest management of gypsy moths in Massachusetts. The problem is to relate the amount of defoliation, measured as above or below a threshold, to egg mass densities in the region. The purpose is to predict the amount of defoliation for the next season for management purposes. In Section 3 we introduce a multivariate probit model used for the analysis and discuss the interpretation of model parameters. In Section 4 we detail the method of composite likelihood as applied to this model. In addition to point es-

Patrick J. Heagerty is Assistant Professor, Department of Biostatistics, University of Washington, Seattle, WA 98195. Subhash R. Lele is Associate Professor, Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205. The authors are grateful to Andrew Liebhold for providing the gypsy moth data and for patiently answering questions. They also acknowledge three anonymous referees whose thorough reviews led to a greatly improved manuscript. Subhash Lele's work was partially supported by Department of Energy grant DE-FC07-94ID73317 to Goodman, Department of Biology, Montana State University, Bozeman.

Journal of the American Statistical Association
September 1998, Vol. 93, No. 443, Theory and Methods

timation, we address the estimation of standard errors and propose a resampling based method. Our predictions, described in Section 5, are empirical Bayes estimates obtained through an approach similar to iterated conditional modes (ICM) described by Besag (1986). In Section 6 we illustrate the method through analysis of the defoliation data. We conclude in Section 7 with a brief discussion. Proofs of consistency and asymptotic normality are provided in the Appendix.

2. DEFOLIATION AND GYPSY MOTH PEST MANAGEMENT

The gypsy moth, *Lymantria dispar (L.)* is probably the most important forest defoliating pest in the northeastern United States. Defoliation and tree mortality associated with gypsy moth outbreaks can cause a multitude of ecological and economic effects (Gottschalk 1993). Every year, more than 250,000 ha of forest land in the United States are sprayed to minimize the adverse effects of defoliation by the gypsy moth (USDA Forest Service 1992).

The Massachusetts Department of Environmental Management monitors gypsy moth defoliation annually in all parts of the state using aerial sketch maps. Maps are sketched during a series of low-level reconnaissance flights in late July when defoliation is at its peak. Thirty percent defoliation is considered the lower threshold for detection from the air. In situations where there is a doubt as to the cause of defoliation, ground checks for the presence of gypsy moth life stages are made. The data are coded as either 0 or 1, depending on whether defoliation exceeded the detection threshold. A grid cell size of 2 km $\times$ 2 km was selected as the standard, and each map consists of 198 $\times$ 93 cells. More details on the data can be obtained from Liebhold et al. (1995).

Hohn, Liebhold, and Gribko (1993) developed a geostatistical model that predicted future defoliation maps from historical defoliation maps. The resulting landscape level predictions of defoliation took into account both the spatial and temporal dynamics. Liebhold et al. (1995) improved upon this modeling by also taking into account the available covariates such as egg mass data through marginal logistic regression. Both of these papers utilize the classical indicator kriging approach (Cressie 1991, pp. 281–283). In this article we apply a model that can be considered "universal indicator cokriging" where not only spatial correlations are modeled but also covariates are used to model the mean response. We utilize a composite likelihood method for estimation of the hierarchical model parameters and use a version of Besag's ICM method for prediction.

3. MODEL

Generalized linear mixed models (Breslow and Clayton 1993; Diggle et al. 1997) are attractive models for dependent data. These models permit regression modeling of the dependence of a response variable on measured covariates while accounting for unmeasured or "random" effects. By adopting a probability model for the random effects, both appropriate inference for regression parameters and prediction of unobserved variables are possible. In the spatial data context, hierarchical model specification first involves identifying the conditional distribution of the response variable $Y(s)$, measured at site s, given both measured covariates $\mathbf{X}(s)$ and unobserved spatially varying random effects, $Z^*(s)$. Typically, the response variables are assumed to be conditionally independent exponential family random variables and a generalized linear model for $E[Y(s)|\mathbf{X}(s), Z^*(s)]$ is formulated. The likelihood function for the observed data, $Y(s)$, is obtained by specifying a second level in the hierarchical model, the joint distribution for the spatial process $Z^*(s)$. Model choices for spatial variation in generalized linear mixed models have included both a conditionally specified Gaussian Markov random field model (e. g., Breslow and Clayton 1993) and an unconditionally specified Gaussian random field model (see Diggle et al. 1997).

The model that we propose for binary spatial data can be considered a special case of the hierarchical generalized linear model discussed by Diggle et al. (1997). But we adopt a *marginal* specification of the mean regression parameter that is made feasible by using a probit link function. In this section we introduce the spatial probit model from a latent variable perspective and discuss parameter identifiability and parameter interpretation.

We assume that spatial binary responses, $Y(s)$, measured at sites $s = 1, 2, \ldots, N$, arise via a threshold model that includes variance components due to spatial dependence and measurement error. Specifically, the threshold model assumes that there exists a Gaussian spatial process, $\mathbf{Z}^*$, with binary responses, $Y(s)$, that are indicators of whether $Z^*(s)$, measured with error, exceeds a certain value

$$Y(s) = \mathbf{1}(Z^*(s) + \varepsilon^*(s) > c),$$

where $\varepsilon^*(s)$ is iid measurement error. Also, $\varepsilon^*(s)$ is independent of $\mathbf{Z}^*$. By specifying the distributions for $Z^*(s)$ and $\varepsilon^*(s)$ as $Z^*(s) \sim \mathrm{N}(\mathbf{X}(s)\boldsymbol{\beta}^*, \tilde{\sigma}^2)$ and $\varepsilon^*(s) \sim \mathrm{N}(0, \tilde{\tau}^2)$, we obtain a probit model for $Y(s)$,

$$P[Y(s) = 1] = \Phi\left(\frac{\mathbf{X}(s)\boldsymbol{\beta}^* - c}{\nu}\right),$$

where $\nu = \sqrt{\tilde{\sigma}^2 + \tilde{\tau}^2}$.

Finally, we obtain a multivariate probit model by specifying the covariance between $Z^*(s)$ and $Z^*(t)$. We assume an isotropic covariance function, $\mathrm{cov}[Z^*(s), Z^*(t)] = \tilde{\sigma}^2\rho(\|s-t\|_2)$.

The complete specification of $\mathbf{Z}^*$ and ε^* yields the multivariate distribution for $\mathbf{Y}$. Of particular interest for our proposed estimation procedure is the pairwise distribution of $Y(s)$ and $Y(t)$. The pairwise distribution can be specified through the univariate marginal probabilities, $P[Y(s) = 1], P[Y(t) = 1]$, and the pairwise marginal probability $P[Y(s) = 1, Y(t) = 1]$. Using the multivariate probit model, we obtain

$$P[Y(s) = 1, Y(t) = 1] = \Phi_2\left(\frac{\mathbf{X}(s)\boldsymbol{\beta}^* - c}{\nu}, \frac{\mathbf{X}(t)\boldsymbol{\beta}^* - c}{\nu}, \frac{\tilde{\sigma}^2\rho(\|s-t\|_2)}{\nu^2}\right),$$

where Φ_2 represents the standardized ($\sigma_1^2 = \sigma_2^2 = 1.0$) bivariate Gaussian distribution function.

We note that only the parameters $c/\nu, \boldsymbol{\beta}^*/\nu$, and $[\tilde{\sigma}^2\rho(\|t-s\|_2)]/\nu^2$ are estimable from the binary data $Y(s)$, because the latent scale is arbitrary. For identifiability and simplicity, we set the total variance $\tilde{\sigma}^2 + \tilde{\tau}^2$ equal to 1. We reparameterize in terms of $Z(s) = (Z^*(s) - c)/\nu, \beta_0 = (\beta_0^* - c)/\nu, \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*/\nu, \varepsilon(s) = \varepsilon^*(s)/\nu$, and $\sigma^2 = \tilde{\sigma}^2/(\tilde{\sigma}^2 + \tilde{\tau}^2)$. Thus the parameter $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1)$ captures the dependence of the unconditional probability $P[Y(s) = 1]$ on covariates $\mathbf{X}(s)$, whereas the proportion of total variation attributable to spatial variation is given by the parameter σ^2. (The proportion due to measurement error is given by $1 - \sigma^2$.) The correlation function $\rho(\|s-t\|_2)$ describes the dependence between sites as a function of distance. For a fixed pair of sites, this correlation is known as the tetrachoric correlation (Pearson 1901).

Given this reparameterization, we can represent the model in terms of the covariate effects, $\mathbf{X}(s)\boldsymbol{\beta}$, spatially correlated deviations, $Z(s) - \mathbf{X}(s)\boldsymbol{\beta}$, and measurement error, $\varepsilon(s)$, as follows:

$$Y(s) = \mathbf{1}(Z(s) + \varepsilon(s) > 0),$$

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\sigma^2, \rho)),$$

and

$$\varepsilon(s) \sim \mathrm{N}(0, (1-\sigma^2)) \quad \text{independent},$$

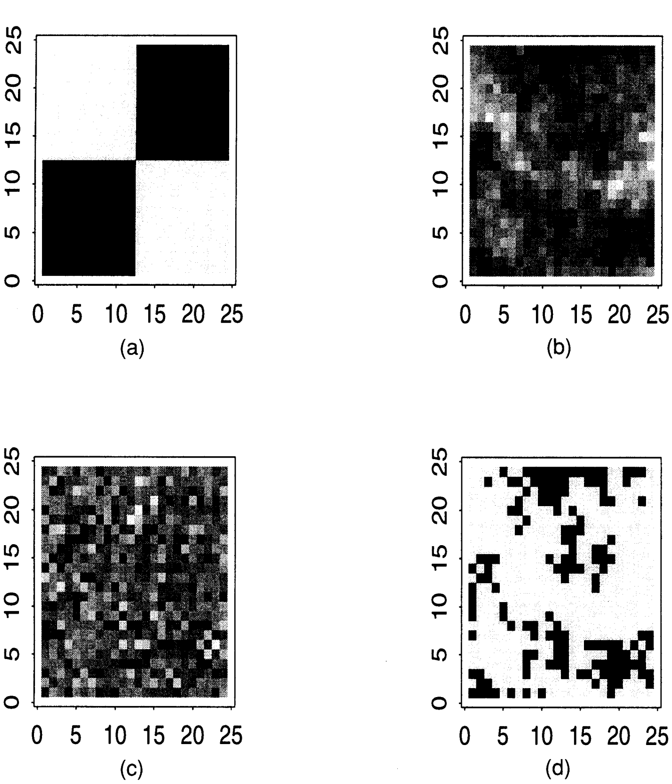

Figure 1. Components of the Spatial Probit Model. Panel (a) shows a hypothetical binary covariate. Panel (b) shows simulated spatially dependent data. Panel (c) shows independent measurement error. Finally, panel (d) shows binary data resulting from a threshold model after combining the components shown in panels (a)–(c).

where $\mathrm{var}[Z(s)] = \sigma^2$ and $\mathrm{cov}[Z(s), Z(t)] = \sigma^2\rho(\|t-s\|_2)$. The process $Z(s)$ can be decomposed into $\mathbf{X}(s)\boldsymbol{\beta} + B(s)$, where $\mathbf{B}$ represents a mean 0 spatially correlated Gaussian process. This notation is similar to that used in the generalized linear mixed model literature (see Breslow and Clayton 1993).

Finally, note that because a probit model is adopted, the parameters $\boldsymbol{\beta}$ and σ^2 also are sufficient to obtain the univariate conditional distribution of $Y(s)$ given $Z(s)$ as follows:

$$P[Y(s) = 1 | Z(s)] = \Phi\left(\frac{Z(s)}{\sqrt{1-\sigma^2}}\right).$$

Figure 1 gives a graphical example of the model components that we are using. This model has several key features. First, it may be used for lattice and nonlattice data, because no assumptions are required about the configuration of the locations, s. The derivation of the model assumes that the binary data are obtained by sampling a continuous spatial region, say A, at a discrete set of points. But it is also legitimate to use the model for data in which the binary responses derive from a regular partition of A. For the example in Section 6, $Y(s)$ represents a measurement over a 2 km × 2 km region, and model parameters are to be interpreted with respect to this areal scale. Also, the parameters $\boldsymbol{\beta}$ capture the *unconditional* dependence of $P[Y(s) = 1]$ on covariates. This interpretation is similar to the parameter interpretation in marginal models (Zeger, Liang, and Albert 1988; see Diggle et al. 1997 for discussion in the spatial context). We make no explicit restrictions on the form of the spatial covariance model used and allow variation due to measurement error. Finally, the hierarchical structure allows simple construction of a valid joint distribution and permits the calculation of empirical Bayes estimates of the spatial process and thus permits kriging with binary data, which we discuss in Section 5.

4. ESTIMATION

Although the complete multivariate distribution of $Y(s)$ is specified by the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \rho)$, likelihood evaluation is computationally impractical. Nonetheless, *pairwise* likelihoods are easily evaluated and can be used to form the kernel of an unbiased estimating function.

4.1 Composite Likelihood Estimation

A composite likelihood is formed by adding together individual component log-likelihoods each of which is a valid marginal or conditional log-likelihood (Lindsay 1988). The key utility of the composite log-likelihood is that the composite score equations form an additive estimating function that can be used to provide consistent parameter estimates in settings where a full maximum likelihood estimator (MLE) is not feasible or is not available. Key examples of successful composite likelihood approaches include working independence generalized estimating equations for longitudinal data (Liang and Zeger 1986) and pseudolikelihood methods for spatial data (Besag 1974). Composite likelihood simply refers to the pooling of likelihood contributions in an additive fashion in circumstances where the components do not necessarily represent independent replicates. Our use of composite likelihood methods allows a high-dimensional likelihood to be approximated by a sum of easily evaluated lower-dimensional components. One advantage of using the composite likelihood score as an estimating function is the ability to evaluate an objective function, the maximized composite likelihood. This facilitates monitoring convergence, assessing multiple roots (if they exist), and displaying of confidence regions by using a rescaled profile composite likelihood function.

4.2 Spatial Probit Pairwise Composite Likelihood Estimation

Consider a single pair of sites s and t. The log-likelihood for the pair of binary responses can be represented in canonical form as

$$\begin{aligned}\log P[Y(s), Y(t)] \\ &= \alpha_0(s,t) + \alpha_1(s,t)Y(s) + \alpha_2(s,t)Y(t) \\ &\quad + \alpha_3(s,t)Y(s)Y(t).\end{aligned}$$

Thus the pairwise score equations are simple quadratic functions of $Y(s)$ and $Y(t)$ given by

$$\mathbf{U}_{(s,t)}(\boldsymbol{\theta}) = \mathbf{D}_{(s,t)}^T \mathbf{V}_{(s,t)}^{-1} \mathbf{R}_{(s,t)},$$

where $\mu(s) = E[Y(s)], \mu(t) = E[Y(t)]$, $\sigma(s,t) = E[(Y(s) - \mu(s))(Y(t) - \mu(t))]$, $\mathbf{D}_{(s,t)} = (\partial/\partial\boldsymbol{\theta})[\mu(s),\ \mu(t),\ \sigma(s,t)]$, $\mathbf{R}_{(s,t)} = \{Y(s) - \mu(s), Y(t) - \mu(t), [Y(s) - \mu(s)][Y(t) - \mu(t)] - \sigma(s,t)\}$, and $\mathbf{V}_{(s,t)} = \operatorname{var}[\mathbf{R}_{(s,t)}]$. See Appendix A for details.

For N spatial observations, we define an estimating function (a composite score function) for the parameter $\boldsymbol{\theta}$ by pooling all possible pairwise score functions as

$$\mathbf{U}_N(\boldsymbol{\theta}) = \frac{1}{W_N} \sum_{(s,t)} w_{(s,t)} \mathbf{U}_{(s,t)}(\boldsymbol{\theta}),$$

where $w_{(s,t)}$ specifies a weight given to the contribution from pair (s,t) and $W_N = \sum w_{(s,t)}$. A simple weighting choice is to let $w_{(s,t)} = 0$ for any pair whose distance exceeds a specified value, and to let $w_{(s,t)} = 1$ otherwise, allowing "distant" pairs to be excluded from the estimating function $\mathbf{U}_N(\boldsymbol{\theta})$. In practice, the choice of $w_{(s,t)}$ can be guided by a preliminary inspection of the correlation between $Y(s)$ and $Y(t)$ as a function of $\|s - t\|_2$ and pairs for which the correlation is small can be excluded ($w_{(s,t)} = 0$) without substantial loss of information regarding σ and ρ but resulting in reduced computational effort.

The estimating function $\mathbf{U}_{(s,t)}(\boldsymbol{\theta})$ is similar to the estimating functions used in GEE2 for longitudinal data (Liang, Zeger, and Qaqish 1992). As such, an important caveat is that the consistency of the estimate of the mean regression parameter, $\hat{\beta}$, may depend on correct specification of the dependence structure. Alternative estimating equation approaches have been proposed in the longitudinal data literature that use paired but separate estimating functions allowing consistent mean parameter estimation even under dependence model violation (Carey, Zeger, and Diggle 1993; Lipsitz, Laird, and Harrington 1991; Prentice 1988). Similar approaches could be adopted for spatial categorical data offering related alternatives to the composite likelihood approach.

4.3 Penalized Composite Likelihood

One modification of the composite score is given by adding a penalty term,

$$\mathbf{U}_N^*(\boldsymbol{\theta}) = \mathbf{U}_N(\boldsymbol{\theta}) - \frac{1}{W_N} \boldsymbol{\Lambda}\boldsymbol{\theta},$$

where $\boldsymbol{\Lambda}$ is a penalty matrix. The penalized composite score can be derived by using the composite likelihood function in conjunction with a Gaussian prior on the parameters $\boldsymbol{\theta}$ where the prior mean is 0 and the prior variance is $\boldsymbol{\Lambda}^{-1}$. (See Heagerty and Zeger 1996 for details in the longitudinal data context.) Our motivation for using the penalization term is simply to offer an extension that may result in stabilized variance component estimates for application with small to moderate sample sizes. For example, in the spatial model where $\operatorname{cov}[Z(s), Z(t)] = \sigma^2 \rho^{\|s-t\|_2}$, defining the variance components in terms of parameter transformations logit(σ) and logit(ρ) results in unrestricted parameter spaces, and the penalty term $\boldsymbol{\Lambda}\theta$, for a positive definite penalty, will generally constrain estimates to be in the interior of the sample space. Therefore, we consider using a nonzero penalty $\boldsymbol{\Lambda}$ as a method for increasing the numerical stability of the estimation algorithm and have in practice used a small penalty, $\boldsymbol{\Lambda} = \lambda\mathbf{I}$, with $0 < \lambda < 1.0$.

4.4 Large-Sample Properties

The key properties of an estimator defined as the root of the function $\mathbf{U}_N(\boldsymbol{\theta}) = \mathbf{0}$ are consistency and asymptotic normality under regularity conditions (Guyon 1995; Lindsay 1988; White 1982).

To detail the large-sample properties, we begin by defining two matrices. The first matrix is the expected derivative of $\mathbf{U}_N(\boldsymbol{\theta})$,

$$\mathcal{I}_0 = \frac{1}{W_N} \sum_{(s,t)} E\left(\frac{\partial}{\partial\boldsymbol{\theta}}\, w_{(s,t)} \mathbf{U}_{(s,t)} \right).$$

The second matrix gives the asymptotic variance of $\hat{\boldsymbol{\theta}}$ as follows assuming that the regularity conditions specified here are satisfied:

$$\mathcal{I}_N^{-1} = (\mathcal{I}_0^{-1}) \left[\frac{1}{W_N^2} \sum_{(s,t),(i,j)} w_{(s,t)} w_{(i,j)} E(\mathbf{U}_{(s,t)} \mathbf{U}_{(i,j)}^T) \right] \times (\mathcal{I}_0^{-1})^T.$$

These matrices allow us to explicitly state the asymptotic properties of the estimator $\hat{\boldsymbol{\theta}}_N$ defined as the root of $\mathbf{U}_N(\boldsymbol{\theta})$.

Notice also that asymptotically as $W_N \to \infty, \mathbf{U}_N$ and $\mathbf{U}_N^*$ (the penalized version) are equivalent, so the large-sample properties of the composite likelihood and penalized composite likelihood estimators are equivalent. For moderate sample sizes, the information matrix

$$\mathcal{I}_N^* = \left(\mathcal{I}_0 + \frac{1}{W_N}\,\mathbf{\Lambda}\right) \times \left[\frac{1}{W_N^2} \sum_{(s,t),(i,j)} w_{(s,t)} w_{(i,j)} E(\mathbf{U}_{(s,t)}\mathbf{U}_{(i,j)}^T)\right]^{-1} \times \left(\mathcal{I}_0 + \frac{1}{W_N}\,\mathbf{\Lambda}\right)^T$$

may be preferable to $\mathcal{I}_N$.

Proposition 1. Define $\hat{\boldsymbol{\theta}}_N$ as the root of the composite score equations $\mathbf{U}_N(\boldsymbol{\theta}) = \mathbf{0}$ for N spatial locations. Let $D_N \subset \Re^d$ be the domain over which $\mathbf{U}_N$ is evaluated and let $|D_N|$ be the cardinality of D_N. Let $\partial S(\boldsymbol{\theta}, \varepsilon)$ denote the boundary of a sphere of radius ε centered at $\boldsymbol{\theta}$. Given regularity conditions on the increasing domain (Guyon 1995, cond. 3.1 and 3.2, p. 108):

A. Assume there exists $\alpha > 0$ and (m_N) a strictly increasing sequence of integers such that

$$\sum_{N \geq 1} N^\alpha |D_{m_N}|^{-1} < \infty,$$

$$\sum_{N \geq 1} N^\alpha \left(\frac{|D_{m_{N+1}} \backslash D_{m_N}|}{|D_{m_N}|}\right)^2 < \infty.$$

and the following conditions on the estimating functions and information (Crowder 1986) hold:

R1. $\mathbf{U}_N(\boldsymbol{\theta})$ is continuous.

R2. $\inf_{\partial S(\boldsymbol{\theta}_{0,\varepsilon})} (\boldsymbol{\theta}_0 - \boldsymbol{\theta})^T E_{\boldsymbol{\theta}_0}[\mathbf{U}_N(\boldsymbol{\theta})] \geq \delta$ for some $\delta > 0$ and N sufficiently large.

R3. $\sup_{\partial S(\boldsymbol{\theta}_{0,\varepsilon})} \|\mathbf{U}_N(\boldsymbol{\theta}) - E_{\boldsymbol{\theta}_0}[\mathbf{U}_N(\boldsymbol{\theta})]\| \to 0$.

Then the composite likelihood estimator is weakly convergent, $\hat{\theta}_N \to \theta_0$.

Proof. See Appendix B.

In addition to consistency, we have asymptotic normality under regularity conditions on the random variables $\mathbf{U}_{(s,t)}$.

Proposition 2. Given that we can apply the central limit theorem to the random vector $\mathbf{U}_{(s,t)}$, and that $\mathbf{U}_N(\boldsymbol{\theta})$ admits a Taylor series expansion, $\hat{\boldsymbol{\theta}}_N$, defined as the solution to the composite score equations, will be asymptotically Gaussian: $\mathcal{I}_N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to \mathrm{N}(\mathbf{0}, \mathbf{I})$.

Proof. See Appendix B.

Table 1. Strong Spatial Dependence; Composite Likelihood Estimates Based on 100 Simulations of 24 × 24 Lattice Data (576 Observations) Using a Radius of 5 Units

Parameter	*Model*	*Estimate (mean)*	*Relative bias*	*Simulation SD*	*Estimated SD (mean)*
β_0	−.50	−.511	−2.2%	.176	.132
β_1	.75	.755	.7%	.110	.091
σ^2	.80	.855	6.9%	1.056	1.228
ρ	.60	.566	−5.7%	.443	.363

NOTE: The standard error calculations are based on transformed dependence parameters, logit(σ) and logit(ρ).

4.5 Standard Error Estimation

Although $E[\mathbf{U}_{(s,t)}\mathbf{U}_{(i,j)}^T]$ can be obtained analytically and numerically using multivariate Gaussian quadrant probabilities, each computation is relatively costly, and the total number of such calculations may be prohibitive. To illustrate the potential difficulty, consider N spatial sites and a composite likelihood based on all $M = \binom{N}{2}$ possible pairs. The number of terms in the expression for the information will be $M + \binom{M}{2}$. Thus although the number calculations for point estimation will be $O(N^2)$, the number of covariance calculations will be $O(N^4)$, which quickly becomes computationally impractical.

When direct evaluation of the covariance is not practical, we suggest using resampling methods similar to those of Carlstein (1986), Künsch (1989), or Sherman (1996). Specifically, we assume that asymptotically

$$N \times E[\mathbf{U}_N(\boldsymbol{\theta})\mathbf{U}_N(\boldsymbol{\theta})^T] \to \mathbf{\Sigma}_\infty.$$

Given N spatial locations we can estimate $\mathbf{\Sigma}_\infty$ using composite score evaluations over K subregions of sizes S_j using

$$\hat{\mathbf{\Sigma}}_\infty = \frac{1}{K}\sum_{j=1}^{K} S_j \mathbf{U}_{S_j}(\hat{\boldsymbol{\theta}})\mathbf{U}_{S_j}(\hat{\boldsymbol{\theta}})^T.$$

The large-sample variance of $\hat{\boldsymbol{\theta}}$ can then be estimated using $\hat{\mathcal{I}}_N^{-1} = \mathcal{I}_0^{-1}[(1/N)\hat{\mathbf{\Sigma}}_\infty]\mathcal{I}_0^{-1}$. This standard error estimator is a special case of a general empirical variance method presented by Lumley and Heagerty (1998).

4.6 Simulation Results

To assess the finite-sample properties of the composite likelihood estimator, we conducted a series of simulation exercises using 24 × 24 lattice data generated using the threshold model. A single covariate, $X_1(s)$, was generated for each site as a realization of a uniform $[-1, 1]$ variable. Let $\mathbf{LL}^T$ be the Cholesky decomposition of the spatial covariance matrix $\mathbf{\Sigma}(\sigma^2, \rho)$. Define the random vector $\mathbf{Z} = \mathbf{X}\beta + \mathbf{Lw}$, where $\mathbf{w}$ is a vector of iid standard normal random variables. Finally, the binary data, $Y(s)$, are the indicators $\mathbf{1}(Z(s) + \varepsilon(s) > 0)$, where $\varepsilon(s)$ represents iid normal random variables with variance $1 - \sigma^2$.

Presented next are the results of two such simulation exercises that illustrate the performance in moderately small datasets under different dependence models. Table 1 is based on realizations from a fairly strong spatially dependent process, ($\sigma^2 = .8, \rho = .6$), whereas Table 2 is based on realizations from a weakly dependent process, ($\sigma^2 = .6, \rho = .4$). Variance estimates were obtained by evaluating the composite score over all possible overlapping 10 × 10 subregions similar to the approach proposed by Sherman (1996). Guidance concerning the selection of the subregion size for estimators defined through estimating functions has been given by Heagerty and Lumley (1998).

Table 2. Weak Spatial Dependence: Composite Likelihood Estimates Based on 100 Simulations of 24 × 24 Lattice Data (576 Observations) Using a Radius of 5 Units

Parameter	Model	Estimate (mean)	Relative bias	Simulation SD	Estimated SD (mean)
β_0	−.50	−.487	2.6%	.095	.082
β_1	.75	.766	2.1%	.106	.093
σ^2	.60	.688	14.7%	1.009	1.061
ρ	.40	.355	−11.3%	.594	.523

NOTE: The standard error calculations are based on transformed dependence parameters, logit(σ) and logit(ρ).

These simulations indicate that the composite likelihood approach can estimate both mean parameters and covariance parameters even for moderate sample sizes. We find little bias in the estimation of the mean parameters, but the covariance parameters tend to overestimate σ and underestimate ρ by as much as 15% in these simulations. Note that the covariance parameter estimates are negatively correlated and yield estimated covariances that are slightly small. For example, the average covariance for sites such that $\|t - s\|_2 = 1$ is estimated for the strong dependence simulations, on average, as .450 ($\sigma^2\rho^1 = .48$) and is estimated for the weak dependence simulations, on average, as .223 ($\sigma^2\rho^1 = .24$).

Simulations reported here indicate that whereas point estimation is adequate for the 24 × 24 lattice data, the standard error estimates tend to slightly underestimate the sampling variability. We expect the standard error estimation to improve with larger sample sizes. This expectation is consistent with the improvement seen in the weak dependence simulations relative to the strong dependence simulations.

5. PREDICTION

One of the major purposes of kriging is prediction. Given the data $\mathbf{Y}$, and the parameter estimates $\hat{\boldsymbol{\theta}}$, we can recover an empirical Bayes estimate of the underlying process $\mathbf{Z}$. We can also use the estimated spatial process to compute a smooth estimate of the probability, $P[Y(s) = 1|Z(s)]$, by using $\Phi[(\hat{Z}(s)/\sqrt{1-\hat{\sigma}^2})]$. One can also obtain a predicted value of the underlying process at a new location. This can be done using the classical kriging predictor (Cressie 1991, pp. 119–123) applied to the empirical Bayes estimate of the underlying process. In this article, we concentrate on the first two possibilities.

Consider first the empirical Bayes prediction of the underlying continuous process $\mathbf{Z}$. This will usually be obtained by considering the full posterior $P[\mathbf{Z}|\mathbf{Y}, \hat{\boldsymbol{\theta}}]$. But the exact evaluation of the full posterior requires construction and inversion of the covariance matrix $\text{cov}(\mathbf{Z}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$, which for sample sizes larger than 500 is computationally prohibitive. We propose two procedures for obtaining posterior modes in large samples. First, we define a local posterior mode (LPM) estimator that uses only data in a local region around the point at which the posterior is being computed. This LPM estimator is naturally influenced by the choice of the local region. To reduce this influence, given an initial estimate of $\mathbf{Z}$, we use an iterative method similar to ICM (Besag 1986). But we again restrict attention to local conditional updating, referring to this method as iterative local conditional modes (ILCM). This procedure is particularly useful after LPM, because it tends to smooth the predictive surface in regions that may be impacted by the choice of the local regions used in LPM. Both procedures can be implemented quickly and easily to construct spatially smoothed empirical Bayes predictive surfaces on either the scale of $\mathbf{Z}$ or the probability scale. We provide technical details of the posterior score equations in Appendix C.

5.1 Local Posterior Modes

Let us consider a single site, s, where we want to obtain the value of $Z(s)$. Consider a neighborhood of s that includes all the sites t such that $\|s - t\|_2 < R$. Let us denote this set of sites by $\mathbf{Z}_s^R$. Consider the local posterior distribution, $[\mathbf{Z}_s^R|\mathbf{Y}_s^R]$. In practice, we assume that the correlations decay quickly as the distances between sites increases. Thus at site s the mode of this local posterior can be a reasonable approximation of the mode of full posterior for this site. The choice of R is governed by balancing competing practical considerations. Computational considerations favor a small value for R, whereas approximation accuracy favors a large value for R.

We now obtain the empirical Bayes estimate for this subregion by maximizing the local posterior. As shown in Appendix C, the local posterior mode $\hat{\mathbf{Z}}_s^R$ is given by solving the local posterior score equations

$$\left[\frac{\partial \boldsymbol{\mu}_s^R}{\partial \mathbf{Z}_s^R}\right]^T (\mathbf{V}_s^R)^{-1}(\mathbf{Y}_s^R - \boldsymbol{\mu}_s^R) - (\boldsymbol{\Sigma}_s^R)^{-1}(\mathbf{Z}_s^R - X_s^R\beta) = \mathbf{0},$$

where the notation $\mathbf{Z}_s^R$ refers to the values $Z(t)$ within R units of site $s, \boldsymbol{\Sigma}_s^R = \text{cov}(\mathbf{Z}_s^R), \mathbf{V}_s^R = \text{var}(\mathbf{Y}_s^R|\mathbf{Z}_s^R)$, and $\boldsymbol{\mu}_s^R = E(\mathbf{Y}_s^R|\mathbf{Z}_s^R)$. Solution to the foregoing equations yields the vector $\hat{\mathbf{Z}}_s^R$. We record only the value $\hat{Z}(s)$. Similar to empirical Bayes estimators in general, the estimator $\hat{Z}(s)$ balances fidelity of the observed data to the latent surface and fidelity of the latent surface to its estimated mean and covariance. The aforementioned process is iterated for all the sites to obtain the initial empirical Bayes estimate of the underlying process.

5.2 Iterated Local Conditional Posterior Modes

One disadvantage of using LPM is the potential for a sharp change in the prediction surface due to the choice of R. One potential solution would be to use LPM but with a larger radius, $R' > R$. But the computational burden increases as R increases. A second approach is to follow LPM with ILCM estimation. ILCM updates the estimated posterior mode at site s by maximizing the posterior mode of $Z(s)$ given the data $\mathbf{Y}$ and the current local empirical Bayes estimates of $\mathbf{Z}$ excluding $Z(s)$ in a region defined by the radius R^*, denoted $\mathbf{Z}_{(-s)}^{R*}$. A single update at a single site is achieved by solving the local conditional posterior score equations

$$\left[\frac{\partial \mu^c(s)}{\partial Z(s)}\right][V^c(s)]^{-1}[Y(s) - \mu^c(s)] - (\sigma_s^{R*})^{-1}\{Z(s) - E[Z(s)|\hat{\mathbf{Z}}_{(-s)}^{R*}]\} = 0,$$

where $\mu^c(s) = E[Y(s)|Z(s)], V^c(s) = [Y(s)|Z(s)]$, and $\sigma_s^{R*} = \text{var}[Z(s)|\mathbf{Z}_{(-s)}^{R*}]$ (see App. C). Note that depending on $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, the conditional posterior of $Z(s)$ may depend on all $Z(t)$ but we use the local conditional posterior as an approximation to the full conditional posterior. Also notice that the posterior depends on $\mathbf{Y}$ only through $Y(s)$. Finally, the choice of R^* here has no relation to the choice of R in LPM.

Besag (1986) proved that iteratively updating the posterior mode estimates in this fashion will converge to the true posterior when the full conditional distribution is used. But his results do not apply to our setting, because we are using a local approximation to the full conditional distribution. In practice we have found that ILCM does converge and can refine the initial estimates given by LPM.

In application, one must choose values for R and R^*. We recommend choosing R and R^* as large as feasible, recognizing that a matrix of size $m \times m$ where $m \approx cR^2$ must be inverted for each location. By using $R < \infty$ or $R^* < \infty$, we are effectively setting partial correlations to 0 for sites greater than $R(R^*)$ units apart. This may impact either the direct calculation of the posterior mode or the iterative calculation based on the conditional expectations $E[Z(s)|Z(t), t \neq s]$. Therefore, we seek R and R^* large enough relative to the range of the dependence so as to minimize the impact of these approximations. In practice, some sensitivity analysis that varies these tuning parameters should be considered.

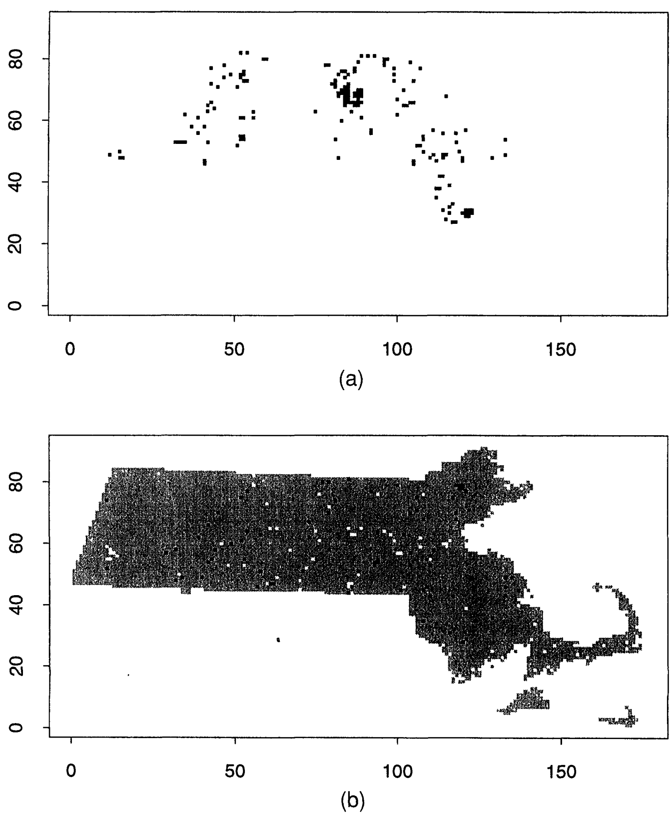

Figure 2. Data for 1990. Panel (a) shows the defoliation sites in black and panel (b) shows the sites that were ground sampled where egg masses were not found (light gray) and where egg masses were found (black).

6. EXAMPLE

In this section we apply the foregoing methodology to analyze data, described in Section 2, on defoliation due to gypsy moths. Scientific interest focuses on three related aspects. The first and foremost is to smooth the threshold data to obtain empirical Bayes predictions of the underlying defoliation risk. In general, the use of covariates should help in this task. Therefore, one also needs to estimate the relationship between various measured covariates and the probability of detectable defoliation. Finally, current and historical data on defoliation can serve as the basis for forecasting future defoliation to guide pest control decision making. We summarize the predictive potential of spatially smooth risk estimates through receiver operating characteristic (ROC) curves calculated using the 1990 risk estimates to predict the 1991 defoliation data.

Our analysis of the gypsy moth data is intended to be illustrative rather than definitive. To truly use the covariates in a fruitful fashion, one needs covariates that have the same coverage as the observed thresholded process. Similarly, to truly address temporal prediction, one should consider using space–time models. For covariates, we have chosen to illustrate with two binary covariates that relate to ground measurements taken only at a subsample of sites. A more definitive and detailed ecological analysis of these data (with additional covariates that have full coverage) and other defoliation data from George Washington National Forest and Shenandoah National Park is presented elsewhere (Lele, Heagerty, and Liebhold 1998). We do, however, illustrate the estimation of regression parameters and the calculation of spatially smooth empirical Bayes estimates and present graphical methods for displaying and quantifying the use of these estimates for predicting 1991 defoliation data.

Figure 2(a) shows the binary defoliation data for 1990, collected via aerial surveillance, and Figure 2(b) shows the sites that also had ground measurements of the presence or absence of gypsy moth egg masses. Only a subset of all sites was chosen to have the ground measurements taken. We use a marginal regression model

$$\Phi^{-1}(P[Y(s) = 1]) = \beta_0 + \beta_1 X_1(s) + \beta_2 X_2(s),$$

where $X_1(s)$ is an indicator variable for a site having ground measurements and $X_2(s)$ is an indicator variable for whether egg masses were found at the site. Because X_2 is nested in X_1, β_2 contrasts the percentage of sites defoliated where egg masses are found versus the percentage of sites defoliated where egg masses are not found. Note that β_2 is based only on contrasts among sampled sites and may not be generalized without further assumptions. Similarly, β_1 compares the defoliation rate at sites with ground samples compared to those sites without ground measurements.

The model parameters were estimated using the penalized composite likelihood approach with a weak penalty, $\boldsymbol{\Lambda} = \lambda \mathbf{I}$ with $\lambda = .1$, although $\lambda = 0$ yields nearly identical results. We formed the composite likelihood using all pairs of sites less than eight units apart, with the choice of the inclusion radius based solely on computational feasibility. The estimated parameters are reported in Table 3. The first model assumes no spatial dependence, as in a standard probit analysis. The naive standard errors correspond to the standard error estimates under the independence model. The next two columns show the standard errors obtained using the spatial bootstrap method discussed in Section 4.5. These standard error estimates show that ignoring the spatial dependence can have serious impact on precision estimates and subsequently on inference. In addition, the estimated standard errors using subregions show the variability in standard error estimates that may result from different subset sizes suggesting the need for research into subset size selection or jackknife estimates (Lele 1991) as an alternative.

We next fit a pair of spatial dependence models using the pairwise composite likelihood method. For these models, we use the covariance model $\text{cov}[Z(s), Z(t)] = \sigma^2 \exp[\log(\rho)\|s-t\|_2^\delta]$ (see Diggle et al. 1997). Our model assumes that the process $\mathbf{Z}$ is isotropic, although alternative anisotropic models may also be considered if suggested by scientific or empirical evidence. We estimated σ and ρ for fixed δ with $\delta = 1.0$ or $\delta = .5$, the latter model allowing a slower correlation decay than the former. In each case the parameter estimates indicate strong spatial dependence. The maximized composite likelihood gives an objective criterion for comparing the two δ values. For these data, $\delta = 1.0$ gives a slightly larger maximized composite likelihood. For the range of distances used to estimate the parameters, $\|s-t\|_2 \leq 8$, a plot of the fitted correlation functions shows the two models to be nearly indistinguishable. But we see that these models yield different correlation decay and thus give different empirical Bayes estimates of $Z(s)$ as seen in Figures 3(b) and 4(a). Note that the posterior contours decay faster in 3(b) than in Figure 4(a).

The coefficient estimate for the presence of egg masses is .561, indicating that the presence of egg masses is a biologically sensible predictor of defoliation risk (although not nominally significant). It should be noted that collecting egg mass data is very difficult, and the strength of the measured relationship can be used to determine whether the costs are warranted. An ideal covariate would be such that it is highly correlated with defoliation and is easily collected or available. The ecology of gypsy moths would suggest that such covariates as that forest type, forest cover, and temperature be considered (A. M. Liebhold, personal communication).

We next obtained empirical Bayes predictions for $Z(s)$ using LPM with a radius of 6 followed by ILCM with a radius of 4. We used these values, $R = 6$ and $R^* = 4$, to have reasonably fast computation and did not find evidence that larger values had appreciable impact on the final estimates. We interpret $\hat{Z}(s)$ as an estimate of the "risk of defoliation" based on the measured covariates and the model for spatial correlation. In describing data collection, we recognized that $Y(s)$ is indeed a thresholded measurement with $Y(s) = 1$ if the fraction of defoliation at a site, $D(s)$, exceeds 30%. Without some actual collection of data on $D(s)$, we can not calibrate $\hat{Z}(s)$ in terms of the unobserved defoliation fraction. We can, however, use $\hat{Z}(s)$ as a standardized score for the sites and/or use it to compute the probability of observable defoliation given covariates and the spatial process $Z(s)$. If scientific interest is in predicting $D(s)$, then composite likelihood methods can be developed to accommodate data $(Y(s), D(s'))$, where the broad collection of binary response data is augmented by selected measurements of the continuous response.

We compared the model-based predictions assuming that $\delta = .5$ and $\delta = 1.0$ with the simple nonparametric approach of a probit transformed locally weighted average of the observed defoliation data where the weights are inversely proportional to distance. The model-based and ad hoc methods can provide very similar smoothing when the "weighting" of data is comparable, as is seen in Figure 4. This is satisfying and gives the empirical Bayes estimates intuitive appeal. However, the method of empirical Bayes has several advantages over the simple statistics. First, the model-based approach can easily accommodate covariate information which may be strongly associated with the response. Second, the "weighting" of proximal observations toward prediction at a given site is based on estimated model pa-

Table 3. Regression Estimates for the 1990 Gypsy Moth Defoliation Data

Independence				
Variable	*Estimate*	*SE naive*	*SE (10 × 13)*	*SE (20 × 26)*
Intercept	−1.922	(.034)	(.090)	(.106)
Sample	−.225	(.392)	(.362)	(.278)
Eggs	.607	(.448)	(.447)	(.450)

Spatial

Variable	*Estimate*	*SE (10 × 13)*	*SE (20 × 26)*	*Estimate*	*SE (10 × 13)*	*SE (20 × 26)*
Intercept	−1.878	(.091)	(.121)	−1.878	(.083)	(.105)
Sample	−.197	(.325)	(.286)	−.197	(.288)	(.236)
Eggs	.561	(.415)	(.466)	.561	(.371)	(.399)
σ	.841	(.818)	(.575)	.995	(1.741)	(1.709)
ρ	.798	(.241)	(.214)	.525	(.256)	(.252)
δ	1.0			0.5		
Max log CL	−136,832.5			−136,843.5		

NOTE: Standard error estimates for the covariance parameters are for the logit transformed parameters. The sizes of the subregions used to compute standard errors were (10 × 13) and (20 × 26). Independence refers to the usual probit point estimate, spatial refers to estimates based on the pairwise composite likelihood.

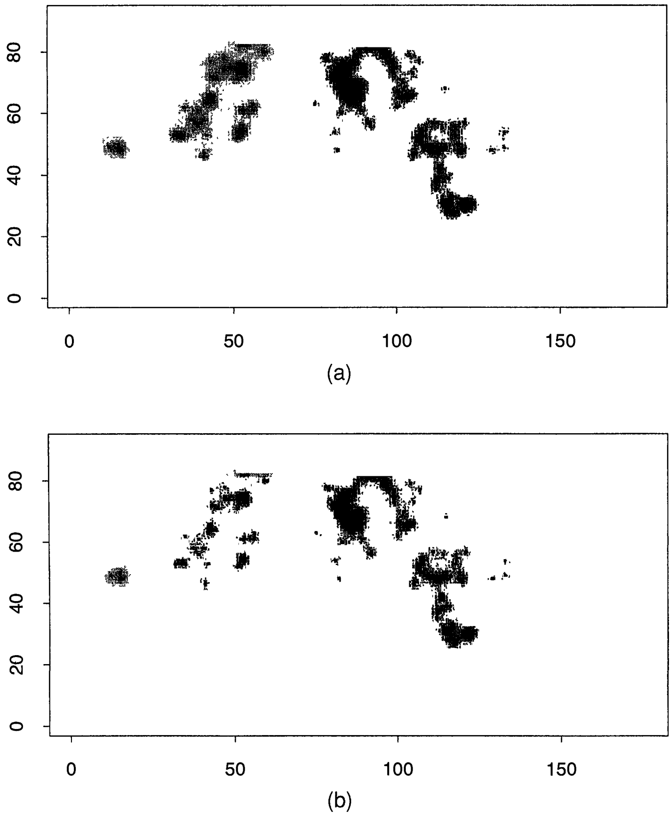

Figure 3. Posterior Estimates of the Spatially Correlated Deviations $Z(s) - \mathbf{X}(s)\beta$ for 1990 Data Using $\delta = 1$. (a) LPM posterior mode; (b) LPM + ILCM posterior mode.

rameters that describe spatial dependence and thus are determined in an objective fashion within the model class used. Furthermore, the methods that we propose can use general covariance structures including anisotropic models. Finally, empirical Bayes methods can, in principle, provide prediction intervals through the use of local posteriors that are available given the estimated parameter values.

An interesting distinction between the two fitted spatial covariance models arises since using $\delta = 1.0$ yields $\hat{\sigma}^2 = .707$, whereas using $\delta = .5$ yields $\hat{\sigma}^2 = .989$. In both cases this parameter represents an extrapolation of the covariance function $\text{cov}[Z(s), Z(t)] = \sigma^2 \exp[\log(\rho)\|s - t\|_2^\delta]$ to $\|s-t\|_2 = 0$, which is not observed in these data. Because the proportion of total variability that can be attributed to pure measurement error is given by $1 - \sigma^2$, we see that the model with $\delta = .5$ admits almost no measurement error. This materializes in the predictions $\hat{Z}(s)$ such that for all 168 sites with $Y(s) = 1$, we have $\hat{Z}(s) > 0$. Recall that we model $Y(s) = \mathbf{1}(Z(s) + \varepsilon(s) > 0)$ and with $\text{var}[\varepsilon(s)] \approx 0$, we can expect $\hat{Z}(s) > 0$ whenever $Y(s) = 1$. This is similar to classical kriging methods where the observed data are interpolated for models that do not incorporate measurement error. On the other hand, for $\delta = 1.0$, the 29% measurement error allows smoothing such that only 39 sites have $\hat{Z}(s) > 0$. Finally, note that these covariance parameters have little impact on the estimated *marginal* regression parameter, $\hat{\beta}$, whereas they would have substantial impact on the estimated *conditional* regression parameters had we chosen to model a conditionally specified parameter.

To compare these prediction approaches in a compact fashion, and because the future observations are binary, we used ROC curves to display the prediction potential. We thresholded the predicted values based on the 1990 data and compared them with the 1991 data. The sensitivity corresponds to the percentage of future defoliation sites that are correctly predicted by $\mathbf{1}(\hat{Z}(s) > c)$ and the specificity to the percentage of future nondefoliated sites correctly predicted by $\mathbf{1}(\hat{Z}(s) \leq c)$. Figure 5 displays dependence of the classification rates on the level c by the ROC curves, indicating the management potential that could be realized by each of the predictors. In this example we find that the model-based empirical Bayes estimates can be used for fairly high correct classification rates, particularly for the model with $\delta = .5$, where correlations decay more slowly as a function of distance.

Figure 6 shows two regions defined by $\mathbf{1}(\hat{Z}(s) > c)$ with c chosen to yield different sensitivity and specificity combinations. Such regions have clear use in providing recommendations for pest control intervention application, given that their predictive use can be validated through further, longer-term studies. The ROC curve also indicates that it is difficult to raise the sensitivity above 93% without a serious reduction in specificity. This reflects a collection of defoliation sites in the NW ($n = 6$ sites) and a collection of defoliation sites in the NE ($n = 10$), comprising approximately 7% of 1991 defoliation sites, that are far from any 1990 defoliation site.

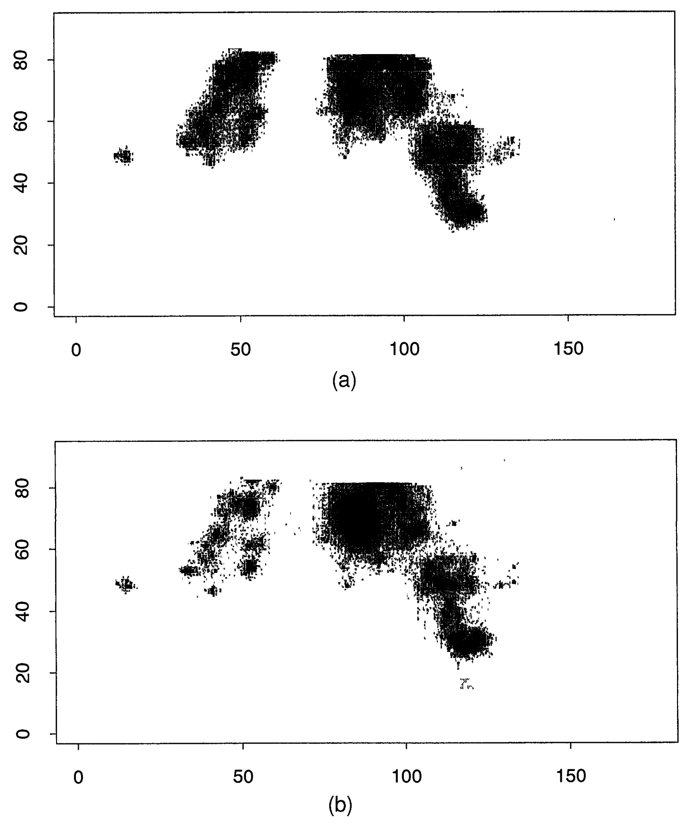

Figure 4. (a) Posterior Estimates of $Z(s)$ for 1990 Using $\delta = .5$; (b) a Plot of Probit Transformed Local Weighted Averages, $\Phi^{-1}[S(s)]$ Where $S(s) = \{1/[\Sigma w(s, t)]\}\Sigma\, w(s, t)Y(t)$, and $w(s, t) = \{\|s - t\|_2 + 1\}^{-2}$.

Finally, we note that the proposed estimation strategy is based on the fit of the model to observed data and not on predictive ability. But the ROC curves indicate that the model can produce spatially smooth estimates that outperform the ad hoc estimators that we considered. As pointed out by a referee, the ad hoc weighted average should be comparable to the empirical Bayes estimates when covariates vary smoothly in space, because the local average combines both spatial dependence and spatial covariate effects. The advantage of the empirical Bayes approach may best be realized in situations where covariates do not have smooth spatial variation.

Unfortunately, the covariance model adopted for the latent Gaussian process can have a major impact on the empirical Bayes estimates. Also, as this example shows, with categorical data it may be difficult to discriminate among nonnested models such as $\delta = 1.0$ versus $\delta = .5$. Our approach has been to assess the direct impact of the covariance form on the ultimate applied use. For prediction of categorical response data we have used ROC curves to characterize and compare competing estimators.

In practice several model choices are required, including the form of the covariance and the selection of tuning parameters. Inspection of the empirical correlations can guide choice of the covariance model. Also, the use of a small penalization term, although asymptotically negligible, can yield improved numerical convergence for small or moderate datasets. If a penalty is used, then analysis of the sensitivity of final results with respect to this choice should be conducted. Finally, for the empirical Bayes estimates, we recommend choosing R and R^* as large as is computationally feasible to have maximal precision in the local approximations to the joint posterior.

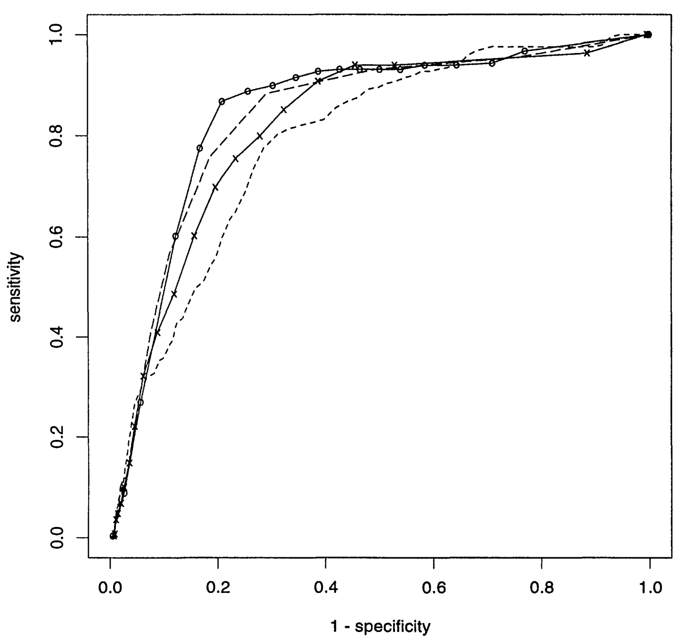

Figure 5. ROC Curves Based on the Posterior Estimates Obtained From the Spatial Probit Models Fit to 1990 Data. The dashed lines are the ROC curves obtained using a weighted mean at each site: $S = \{1/[\Sigma\ w(s, t)]\}\ \Sigma\ w(s, t)Y(t)$, where $w(s, t) = \{\|s - t\|_2 + 1\}^{-p}$ for p = 1, 2. —o— Posterior mode (d = .5); —×—, posterior mode (d = 1.0); --- weighted average (p = 1); – – –, weighted average (p = 2).

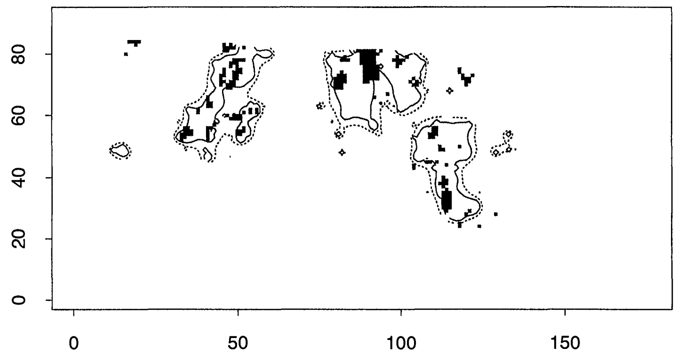

Figure 6. Forecast Contours Based on the Posterior Modes Obtained From the Spatial Probit Model with δ = .5. These define map regions that yield (sensitivity, specificity) combinations of (89%, 72%; – – –) and (78%, 83%; —). The 1991 defoliation data are also shown.

Through this example we have illustrated the use of hierarchical models for the analysis of spatial binary data. The application to ecological pest data shows how covariates can be assessed as correlates to damage as well as how empirical Bayes methods can be employed to create spatially smooth surfaces useful for management or intervention planning. Methods that accommodate space–time data with a goal of accurate prediction need to be developed for binary data and the methods we use here give an indication of future research directions.

7. DISCUSSION

We have proposed a method for regression estimation and prediction for spatial categorical data. By using a hierarchical model that yields easily evaluated marginal first and second moments, we are able to use composite likelihood methods (Lindsay 1988) for estimation and propose empirical Bayes methods similar to those of Besag (1986) for prediction.

Indeed, there are other approaches to estimation and prediction for spatial categorical data. Albert and McShane (1995) utilized a generalized estimating equations approach for the estimation of marginal mean parameters. They do not consider a hierarchical structure presumably because prediction is not a key goal of their application.

Diggle et al. (1997) and Waller et al. (1997) utilize a full Bayesian framework, using Markov chain Monte Carlo (MCMC) for estimation and prediction. Key advantages of this approach are that full posterior distributions are available and that posterior estimates also include the uncertainty in parameter estimates, an aspect of the use of empirical Bayes that we do not address. But it is increasingly well known that caveats are associated with MCMC methods. Both Diggle et al. (1997) and Waller et al. (1997) cautioned about parameterization and convergence monitoring. In practical settings the impact of informative priors must be addressed and the use of noninformative priors can be dangerous (Hobert and Casella 1996). In addition, the inherent computational burden of MCMC methods can hinder their application to large datasets of practical importance.

Approximate likelihood methods provide another alternative. These include methods based on a Laplace approximation discussed as penalized quasi-likelihood (PQL) by Breslow and Clayton (1993). These methods have proven adequate for count data; however, the potential for severe bias with binary data is recognized. McCulloch (1997) illustrated problems associated with PQL for binary data. In addition, PQL requires an N^2 matrix inversion for blocks of N dependent observations, limiting its use to moderate-sized spatial datasets.

The composite likelihood approach has its own limitations as well. It is well suited for situations in which lower-dimensional marginal moments are easily described. But if the underlying process is modeled as a non-Gaussian process, or if, for example, a logit link is used, then the use of composite likelihood may require the use of one- and two-dimensional numerical integration.

Spatially correlated binary data occur in many practical situations. Indicator kriging has shown its usefulness in applications that include environmental remediation and ecological pest control. As noted by Albert and McShane (1995), spatial categorical data are also increasingly common in medical applications where response measurements are obtained at different physiologic locations. Application of hierarchical models to dependent categorical data can provide for both regression analysis and empirical Bayes smoothing. In this article we have suggested a method of estimation and prediction that is particularly suited to large datasets where likelihood methods for hierarchical models are not feasible.

APPENDIX A: COMPOSITE SCORE EQUATIONS

The estimator $\hat{\boldsymbol{\theta}}_N$ is defined as the solution to $\mathbf{U}_N(\boldsymbol{\theta}) = \mathbf{0}$, where $\mathbf{U}_N(\boldsymbol{\theta}) = 1/W_N \sum w_{(s,t)}\mathbf{U}_{(s,t)}(\boldsymbol{\theta})$, where each $\mathbf{U}_{(s,t)}$ is the derivative of the log-likelihood based on the single pair of observations (s,t). The parameter $\boldsymbol{\theta}$ is comprised of a marginal mean regression parameter, $\boldsymbol{\beta}$, and a covariance parameter, $\boldsymbol{\gamma}$, that specifies the correlation function $\rho(s,t) = \operatorname{cov}[Z(s)+\varepsilon(s), Z(t)+\varepsilon(t)]$.

The pairwise score is given as

$$\mathbf{U}_{(s,t)}(\boldsymbol{\theta}) = \mathbf{D}_{(s,t)}^T \mathbf{V}_{(s,t)}^{-1} \mathbf{R}_{(s,t)},$$

where $\mu(s) = E[Y(s)] = \Phi_1[\mathbf{X}(s)\boldsymbol{\beta}], \sigma(s,t) = E[(Y(s)-\mu(s))(Y(t)-\mu(t))] = \pi(s,t)-\mu(s)\mu(t), \mathbf{D}_{(s,t)} = (\partial/\partial\boldsymbol{\theta})[\mu(s),\mu(t),\sigma(s,t)], \mathbf{R}_{(s,t)} = \{Y(s)-\mu(s), Y(t)-\mu(t), [Y(s)-\mu(s)][Y(t)-\mu(t)]-\sigma(s,t)\}$, and $\mathbf{V}_{(s,t)} = \operatorname{var}[\mathbf{R}_{(s,t)}]$. Here we denote $\pi(s,t) = P[Y(s)=1, Y(t)=1] = \Phi_2[\mathbf{X}(s)\boldsymbol{\beta}, \mathbf{X}(t)\boldsymbol{\beta}, \rho(s,t)]$, where Φ_2 is the standardized $(\sigma_1^2 = \sigma_2^2)$ bivariate Gaussian distribution function and Φ_1 the univariate distribution function.

The elements of $\mathbf{D}_{(s,t)}$ are

$$\begin{bmatrix} \frac{\partial}{\partial\boldsymbol{\beta}}\mu(s) & 0 \\ \frac{\partial}{\partial\boldsymbol{\beta}}\mu(t) & 0 \\ \frac{\partial}{\partial\boldsymbol{\beta}}[\pi(s,t)-\mu(s)\mu(t)] & \frac{\partial}{\partial\boldsymbol{\gamma}}\pi(s,t) \end{bmatrix},$$

where under the marginal probit model we have $(\partial/\partial\boldsymbol{\beta})\mu(s) = \phi_1([\mathbf{X}(s)\boldsymbol{\beta}])\mathbf{X}(s)$ and $(\partial/\partial\boldsymbol{\beta})\pi(s,t) = \phi_1[\mathbf{X}(s)\boldsymbol{\beta}]\Phi_1[\xi(s,t)]\mathbf{X}(s) + \phi_1[\mathbf{X}(t)\boldsymbol{\beta}]\Phi_1[\xi(t,s)]\mathbf{X}(t)$, where $\xi(s,t) = (\mathbf{X}(s)\boldsymbol{\beta} - \rho(s,t)\mathbf{X}(t)\boldsymbol{\beta})/[\sqrt{1-\rho(s,t)^2}]$. Also, $(\partial/\partial\boldsymbol{\gamma})\pi(s,t) = \phi_2[\mathbf{X}(s)\boldsymbol{\beta}, \mathbf{X}(t)\boldsymbol{\beta}, \rho(s,t)]([\partial\rho(s,t)]/\partial\boldsymbol{\gamma})$. Here ϕ_1 and ϕ_2 are the univariate and bivariate standard normal density functions.

Finally, the elements of the covariance matrix $\mathbf{V}_{(s,t)}$ are

$$\begin{bmatrix} \mu(s)[1-\mu(s)] & \sigma(s,t) & \sigma(s,t)[1-2\mu(s)] \\ \sigma(s,t) & \mu(t)[1-\mu(t)] & \sigma(s,t)[1-2\mu(t)] \\ \sigma(s,t)[1-2\mu(s)] & \sigma(s,t)[1-2\mu(t)] & V_{22}(s,t) \end{bmatrix},$$

where

$$V_{22}(s,t) = \pi(s,t)[1-\pi(s,t)] + \pi(s,t)[6\mu(s)\mu(t) - 2\mu(s) - 2\mu(t)] + \mu(s)\mu(t)[\mu(s)+\mu(t)-4\mu(s)\mu(t)].$$

APPENDIX B: LARGE-SAMPLE PROPERTIES OF SPATIAL PROBIT COMPOSITE LIKELIHOOD ESTIMATOR

This appendix sketches the proof of the consistency and asymptotic normality of the maximum pairwise composite likelihood estimator for spatial binary data.

1. Consistency

Define the following:

1. $Y(s) = 0/1$, a binary indicator for sites $s = 1, 2, \ldots, N$.
2. Let $\|s-t\|_2$ be a distance metric between pairs of sites.
3. Let $\mathbf{U}_{(i,j)}(\boldsymbol{\theta}) = \mathbf{D}_{(i,j)}^T(\boldsymbol{\theta})\mathbf{V}_{(i,j)}^{-1}(\boldsymbol{\theta})\mathbf{R}_{(i,j)}(\boldsymbol{\theta})$ be the score equations based on a single pair of sites (i,j), where $\mathbf{R}_{(i,j)}^T = [Y(i)-\mu(i), Y(j)-\mu(j)$, and $S(i,j)-\sigma(i,j)]$ such that $S(i,j) = [Y(i)-\mu(i)][Y(j)-\mu(j)]$.
4. Let $w_{(i,j)}$ be a bounded scalar weight such that for all $i, w_{(i,j)} > 0$ only for j such that $\|i-j\|_2 < R$, where R is a fixed constant. Let $W_N = \sum w_{(i,j)}$.
5. Define $\mathbf{U}_N(\boldsymbol{\theta}) = (1/W_N)\sum w_{(i,j)}\mathbf{U}_{(i,j)}(\boldsymbol{\theta})$.

The spatial probit composite likelihood estimator is consistent and asymptotically Gaussian under regularity conditions given by Crowder (1986) or Guyon (1995). We restate those regularity conditions, make further simplifying assumptions, and show that the required conditions are satisfied by the spatial probit model.

In either case we assume that the spatial domain is increasing in a regular fashion. Let $D_N \subset \Re^d$ be the domain over which U_N is evaluated and let $|D_N|$ be the cardinality of D_N. The regularity conditions that we assume are given in the statement of Proposition 1 and are denoted conditions A, R1, R2, and R3.

Furthermore, we assume the following simplifying assumptions that are sufficient for R2 and R3:

A1. Assume a spatial covariance model such that $\sup E[\mathbf{U}_{(i,j)}] < \infty$ and $\sup_N E(N\mathbf{U}_N^2) = K < \infty$.

A2. Assume $(1/W_N)\sum w_{(i,j)}\mathbf{D}_{(i,j)}^T\mathbf{V}_{(i,j)}^{-1} \to \mathbf{B}_\infty$, a finite limit.

A3. Define $\mathcal{I}_N^{(1)} = (1/W_N)\sum E[w_{(i,j)}(\partial/\partial\boldsymbol{\theta})\mathbf{U}_{(i,j)}(\boldsymbol{\theta})]$. Assume $\mathcal{I}_N^{(1)} \to \mathcal{I}_\infty^{(1)}$ where λ_1, the minimum eigenvalue of $\mathcal{I}_\infty^{(1)}$, is strictly positive.

A4. Assume that $(\partial/\partial\boldsymbol{\theta})\mathbf{D}_{(i,j)}^T\mathbf{V}_{(i,j)}^{-1}$ is bounded uniformly in (i,j).

Condition A1 is as given in theorem (3.2.1) of Guyon (1995), which yields strong convergence of $\mathbf{U}_N \to \mathbf{0}$. The covariance conditions need to be verified in practice but can be shown to be satisfied by the exponential correlation models we have considered here.

Condition A2, with the continuity assumption R1, allows a first-order Taylor series expansion of $E_{\boldsymbol{\theta}_0}[\mathbf{U}_N(\boldsymbol{\theta})]$ yielding

$$(\boldsymbol{\theta}_0 - \boldsymbol{\theta})^T E_{\boldsymbol{\theta}_0}[\mathbf{U}_N(\boldsymbol{\theta})] = (\boldsymbol{\theta}_0 - \boldsymbol{\theta})^T \mathcal{I}_N^{(1)} (\boldsymbol{\theta}_0 - \boldsymbol{\theta}) + o\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2.$$

This expression and condition A3 are then sufficient for condition R2. Condition A3 for the spatial probit model implies that $(1/W_N)\sum w_{(i,j)}\mathbf{D}_{(i,j)}^T \mathbf{V}_{(i,j)}^{-1}\mathbf{D}_{(i,j)}$ converges to a matrix with positive eigenvalues, which is satisfied in practice by assuming that covariates enter into the sample in a regular fashion.

Lemmas 2.2 and 3.2 of Crowder (1986) show that condition R3 is equivalent to conditions on $(\partial/\partial\boldsymbol{\theta})\mathbf{U}_N(\boldsymbol{\theta})$. Assumption A4, together with the strong law of large numbers for $\mathbf{U}_N$ is then a sufficient condition for R3 and can be verified in practice by assuming both bounded covariates and a bounded parameter space.

2. Asymptotical Normality

Guyon (1995) gave quite general conditions for the asymptotic normality of estimators defined through a "contrast function," the composite likelihood being a special case. We restate a slightly stronger version of Guyon's regularity conditions in our notation and discuss verification for our estimator.

Consider the following conditions:

H1. There exists an open neighborhood V of $\boldsymbol{\theta}_0 \in \Re^p$ over which $\mathbf{U}_N$ is continuously differentiable, and there exists an integrable random variable h such that for all elements of $(\partial/\partial\boldsymbol{\theta})\mathbf{U}_N(\boldsymbol{\theta})$ and all $\boldsymbol{\alpha} \in V, |(\partial/\partial\boldsymbol{\theta})\mathbf{U}_N(\boldsymbol{\alpha}, \mathbf{Y})| < h(\mathbf{Y})$.

H2. There exists a limiting covariance matrix $\mathcal{I}_\infty^{(2)}$ such that $\mathcal{I}_N^{(2)} = N \times E[\mathbf{U}_N(\boldsymbol{\theta})\mathbf{U}_N(\boldsymbol{\theta})^T]$ where

a. $\mathcal{I}_\infty^{(2)} > 0$ and $\mathcal{I}_N^{(2)} \geq \mathcal{I}_\infty^{(2)}$ for $N \geq m$ for some m and

b. $\sqrt{N}[\mathcal{I}_N^{(2)}]^{-1/2}\mathbf{U}_N \to \mathrm{N}(\mathbf{0}, \mathbf{I}_{p\times p})$.

H3. There exists a sequence of nonstochastic matrices $\mathcal{I}_N^{(1)}$ such that

a. There exists $\mathcal{I}_\infty^{(1)}$ such that $\mathcal{I}_N^{(1)} > \mathcal{I}_\infty^{(1)}$ for $N \geq m$ for some m and

b. $\lim_N[(\partial/\partial\boldsymbol{\theta})\mathbf{U}_N - \mathcal{I}_N^{(1)}] = \mathbf{0}$ in probability.

Given that these conditions are satisfied, we can use theorem (3.4.5) of Guyon (1995), yielding

$$\sqrt{N}[\mathcal{I}_N^{(2)}]^{-1/2}\mathcal{I}_N^{(1)}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \to N(\mathbf{0}, \mathbf{I}_{p\times p}).$$

Condition H2 requires that a central limit theorem can be applied to the sequence $\mathbf{U}_N$. Standard mixing conditions, such as those given in Theorem (3.3.1) of Guyon (1995), yield the desired result and are satisfied in our application by assuming an exponential correlation decay and a finite radius of $w_{(i,j)} > 0$ for each point. Satisfaction of conditions H1 and H3 is a result of assumptions A3 and A4.

APPENDIX C: POSTERIOR SCORE EQUATIONS

We adopt an empirical Bayes approach to the creation of spatially smooth estimates and thus need to find posterior means or posterior modes. For the calculation of posterior modes, we solve score equations defined as $(\partial/\partial Z)\log(P[\mathbf{Z}|\mathbf{Y}, \hat{\theta}]) = \mathbf{0}$. The log-posterior distribution (for fixed $\boldsymbol{\theta}$), is given by $K(\boldsymbol{\theta}, \mathbf{Y}) + \sum_{s=1}^N \log(P[Y(s)|Z(s), \boldsymbol{\theta}]) - 1/2(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})$, because we assume that the data $Y(s)$ are conditionally independent given $\mathbf{Z}$ and that $\mathbf{Z}$ has a joint Gaussian distribution with mean $\mathbf{X}\boldsymbol{\beta}$ and covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Because we use a generalized linear model for $\mathbf{Y}$ with $\mathbf{Z}$ as the linear predictor, the derivative of the log-likelihood with respect to $Z(s)$ will be $\{[\partial\mu^c(s)]/[\partial Z(s)]\}[V^c(s)]^{-1}[Y(s) - \mu^c(s)]$, where $V^c(s) = \mathrm{var}[Y(s)|Z(s)]$ and $\mu^c(s) = E[Y(s)|Z(s)] = \Phi([Z(s)]/\sqrt{1-\sigma^2})$. Therefore, the derivative of the log posterior will have the form

$$\left[\frac{\partial\boldsymbol{\mu}^c}{\partial\mathbf{Z}}\right]^T \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}^c) - \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}),$$

where $\mathbf{V} = \mathrm{diag}[V^c(s)]$.

Note also that the posterior distribution for $Z(s)$ given all $Z(t)\, t \neq s$ can be obtained as $[Z(s)|Z(t), t \neq s, \mathbf{Y}] \propto [Z(s), \mathbf{Y}|Z(t), t \neq s] = [\mathbf{Y}|\mathbf{Z}][Z(s)|Z(t), t \neq s] \propto [Y(s)|Z(s)][Z(s)| Z(t), t \neq s]$. Thus conditional posterior score equations combine the likelihood contribution from site s only with the Gaussian conditional distribution of $Z(s)$ given other sites, $Z(t), t \neq s$,

$$\left[\frac{\partial\mu^c(s)}{\partial Z(s)}\right] V^c(s)^{-1}[Y(s) - \mu^c(s)] - \sigma_s^{-1}\{Z(s) - E[Z(s)|Z(t), t \neq s]\} = 0,$$

where $\sigma_s = V[Z(s)|Z(t), t \neq s]$.

[Received February 1997. Revised March 1998.]

REFERENCES

Albert, P. S., and McShane, L. M. (1995), "A Generalized Estimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data," *Biometrics*, 51, 627–638.

Besag, J. E. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 192–236.

——— (1986), "On the Statistical Analysis of Dirty Pictures," *Journal of the Royal Statistical Society*, Ser. B, 48, 259–279.

Besag, J. E., York, J., and Mollie, A. (1991), "Bayesian Image Restoration With Applications in Spatial Statistics" (with discussion), *Annals of the Institute of Statistical Mathematics*, 43, 21–59.

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–24.

Carey, V. C., Zeger, S. L., and Diggle, P. J. (1993), "Modelling Multivariate Data With Alternating Logistic Regressions," *Biometrika*, 80, 517–526.

Carlin, B. P., and Louis, T. A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, New York: Chapman and Hall.

Carlstein, E. (1986), "The Use Of Subseries Values for Estimating the Variance of a General Statistic From a Stationary Sequence," *The Annals of Statistics*, 14, 1171–1179.

Clayton, D., and Kaldor, J. (1987), "Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping," *Biometrics*, 43, 671–681.

Cressie, N. (1991), *Statistics for Spatial Data*, New York: Wiley.

Crowder, M. J. (1986), "On Consistency and Inconsistency of Estimating Equations," *Econometric Theory*, 2, 305–330.

Diggle, P. J. (1988), "An Approach to the Analysis of Repeated Measures," *Biometrics*, 44, 959–971.

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

Diggle, P. J., Moyeed, R. A., and Tawn, J. A. (1997), "Model-Based Geostatistics" (with discussion), unpublished manuscript submitted to *Applied Statistics*.

Gottschalk, (1993), "Impacts, Silviculture and the Gypsy Moths," *The Lymantriidae: A Comparison of Features of New and Old World Tussock Moths*, eds. W. E. Walker and K. A. McManus, USDA Forest Service, General Technical Report NE-123.

Guyon, X. (1995), *Random Fields on a Network: Modeling, Statistics, and Applications*, New York: Springer-Verlag.

Haining, R. J. (1989), *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge, U.K.: Cambridge University Press.

Heagerty, P. J., and Lumley, T. S. (1998), "Window Resampling and Estimating Functions," Technical Report 151, University of Washington, Dept. of Biostatistics.

Heagerty, P. J., and Zeger, S. L. (1996), "Marginal Regression Models for Clustered Ordinal Responses," *Journal of the American Statistical Association*, 91, 1024–1034.

Hobert, J. P., and Casella, G. (1996), "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Models," *Journal of the American Statistical Association*, 91, 1461–1473.

Hohn, M., Liebhold, A. E., and Gribko, L. (1993), "Geostatistical Model for Forecasting the Spatial Dynamics of Defoliation Caused by the Gypsy Moth, *Lymantria dispar* (Lepidoptera: Lymantriidae)," *Environmental Entomology*, 22, 1066–1075.

Journel, A. E. (1983), "Nonparametric Estimation of Spatial Distributions," *Journal of the International Association for Mathematical Geology*, 15, 445–468.

Keenan, D. (1982), "A Time Series Analysis of Binary Data," *Journal of the American Statistical Association*, 77, 816–821.

Künsch, H. (1989), "The Jackknife and Bootstrap for General Stationary Observations," *The Annals of Statistics*, 17, 1217–1241.

Lele, S. (1991), "Jackknifing Linear Estimating Equations: Asymptotic Theory and Applications in Stochastic Processes," *Journal of the Royal Statistical Society*, Ser. B, 53, 253–267.

Lele, S., Heagerty, P., and Liebhold, A. (1998), "Spatial Analysis of Defoliation Data Due to Gypsy Moths (*Lepidoptera: Lymantriidae*): An Empirical Bayes Approach," unpublished manuscript in preparation.

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992), "Multivariate Regression Analyses for Categorical Data" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 54, 3–40.

Liebhold, A. M., Elkinton, J. S., Zhou, J., Hohn, M., Rossi, R. E. Boettner, G. H., Boettner, C. W., Burnham, C., and McManus, M. L. (1995), "Regional Correlation of Gypsy Moth (*Lepidotera: Lymantriidae*) Defoliation With Counts of Egg Masses, Pupae and Male Moths," *Environmental Entomology*, 24, 193–203.

Lindsay, B. G. (1988), "Composite Likelihood Methods," *Contemporary Mathematics*, 80, 221–239.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1991), "Generalized Estimating Equations for Correlated Binary Data: Using the Odds Ratio as a Measure of Association," *Biometrika*, 78, 153–160.

Lumley, T. S., and Heagerty, P. J. (1998), "Weighted Empirical Adaptive Variance Estimators for Correlated Data Regression," unpublished manuscript submitted to *Journal of the Royal Statistical Society*, Ser. B.

Matheron, G. (1976) "A Simple Substitute for Conditional Expectation: The Disjunctive Kriging," in *Advanced Geostatistics in the Mining Industry*, eds. M. Guarascio, M. David, and C. Huijbrechts, Dordrecht, Holland: Reidel, pp. 221–236.

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170.

Pearson, K. (1901), "Mathematical Contributions to the Theory of Evolution VII. On the Correlation of Characters Not Quantitatively Measured," *Philisophical Transactions of the Royal Society, London*, Ser. A, 195, 1–47.

Prentice, R. L. (1988), "Correlated Binary Regression With Covariates Specific to Each Binary Observation," *Biometrics*, 44, 1033–1048.

Sherman, M. (1996), "Variance Estimation for Statistics Computed From Spatial Lattice Data," *Journal of the Royal Statistical Society*, Ser. B, 58, 509–523.

U.S.D.A. Forest Service (1992), "Forest Insect and Disease Conditions in the United States 1991," United States Department of Agriculture, Forest Service, Forest Pest Management, Washington D.C.

Waller, L. A., Carlin, B. P., Xio, H., and Gelfand, A. (1997), "Hierarchical Spatio-Temporal Mapping of Disease Rates," *Journal of the American Statistical Association*, 92, 607–617.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–24.

Yasui, Y., and Lele, S. R. (1997), "A Regression Method for Spatial Disease Rates: An Estimating Function Approach," *Journal of the American Statistical Association*, 92, 21–32.

Zeger, S. L., Liang, K.-Y., and Albert, P. (1988), "Models for Longitudinal Data: A Generalized Estimating Equations Approach," *Biometrics*, 44, 1049–1060.