Modeling Spatial-Temporal Binary Data Using Markov Random Fields

Jun ZHU, Hsin-Cheng HUANG, and Jungpin WU

An autologistic regression model consists of a logistic regression of a response variable on explanatory variables and an autoregression on responses at neighboring locations on a lattice. It is a Markov random field with pairwise spatial dependence and is a popular tool for modeling spatial binary responses. In this article, we add a temporal component to the autologistic regression model for spatial-temporal binary data. The spatial-temporal autologistic regression model captures the relationship between a binary response and potential explanatory variables, and adjusts for both spatial dependence and temporal dependence simultaneously by a space-time Markov random field. We estimate the model parameters by maximum pseudo-likelihood and obtain optimal prediction of future responses on the lattice by a Gibbs sampler. For illustration, the method is applied to study the outbreaks of southern pine beetle in North Carolina. We also discuss the generality of our approach for modeling other types of spatial-temporal lattice data.

Key Words: Autologistic model; Gibbs sampler; Markov chain Monte Carlo; Maximum pseudo-likelihood; Spatial-temporal model.

1. INTRODUCTION

The southern pine beetle has caused severe damage to pine forests in the southern states of the United States and hence is of great concern. Research has found that the outbreaks are influenced by factors such as host volumes, physiographic properties of the fields, and seasonal temperature. Further, outbreaks of the southern pine beetle in forests throughout the southern United States show visible spatial and temporal patterns (see, e.g., Mawby and Gold 1984; Bailey 1995). In particular, temporal patterns of the outbreaks have been studied. For example, Pye (1993) reported a cycle of length six to seven years for the outbreaks in the southern United States; Turchin, Lorio, Taylor, and Billings (1991) found temporal autocorrelation at a lag of one to two years for some populations in eastern Texas.

To our knowledge, Gumpertz, Wu, and Pye (2000) were the first to develop a statistical model for southern pine beetle outbreak which accounts for potential explanatory variables

while adjusting for spatial and temporal autocorrelation. They studied the outbreaks of southern pine beetle in 301 counties of three states in the United States (Georgia, North Carolina, and South Carolina) from 1960 to 1996. In this article, we focus our attention on the outbreak data from North Carolina. Aggregated over time, the outbreaks show clear positive spatial correlation (Figure 1); whereas aggregated over 100 counties, the outbreaks show positive temporal dependence (Figure 2). In Gumpertz et al. (2000), a marginal logistic regression model was used (see also Diggle, Liang, and Zeger 1994). Statistical models were constructed by first estimating the temporal dependence for each location and then accounting for spatial dependence among locations. As a consequence, statistical inference, including parameter estimation and response prediction, was performed in a stepwise fashion. Even though the inference was optimal at each step, optimality might not be guaranteed for the final inference. The primary purpose of this article is to develop a spatial-temporal autologistic regression model that would systematically model the relationship between a binary response variable and potential explanatory variables, while accounting for spatial dependence and temporal dependence *simultaneously*.

Our approach will be to add a temporal component to the atemporal autologistic models developed by Besag (1972, 1974). Autologistic models account for spatial dependence among binary variables on a regular or irregular lattice. With the specification of a logistic regression, autologistic regression models can be used to model relationships between the binary response variable and potential explanatory variables, while incorporating spatial correlation (see, e.g., Cressie 1993, sec. 6.5.1). Consider representative sites $s_1, \ldots, s_n$ on a spatial lattice. For a given neighborhood structure, let $N_i \equiv \{j : s_j \text{ is a neighbor of } s_i\}$. For notational convenience, let $j \sim i$ if $j \in N_i$. Neighborhood structures are oftentimes based on proximity among the representative sites. For example, on a regular square lattice, commonly used neighborhoods include first order, diagonal, and second order. Let $Y_1, \ldots, Y_n$ denote binary responses on the lattice, where $Y_i \equiv Y(s_i) = 0$ or 1. The joint distribution of $\mathbf{Y} \equiv (Y_1, \ldots, Y_n)'$ for an autologistic regression model can be formulated in a way similar to Greig, Porteous, and Seheult (1989):

$$f(\mathbf{Y}) \propto \exp\left\{\sum_{i=1}^{n}\sum_{k=0}^{p} \theta_k X_{k,i} Y_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j\sim i} \theta_{ij}[Y_i Y_j + (1 - Y_i)(1 - Y_j)]\right\}, \quad (1.1)$$

where $X_{k,i} \equiv X_k(s_i)$ denotes the kth explanatory variable at site s_i, θ_k denotes the kth logistic regression coefficient corresponding to $X_k(\cdot)$, with $k = 0, \ldots, p$. Further, θ_{ij} denotes the autoregression coefficient between the ith site and the jth site, such that $\theta_{ij} = \theta_{ji}$ and $\theta_{ij} \neq 0$ only if $j \sim i$. It follows from (1.1) that the distribution of Y_i conditional on all other Y_j, denoted as $f(Y_i|\mathbf{Y} \setminus Y_i)$, depends only on those at the neighboring sites:

$$\begin{aligned} f(Y_i|\mathbf{Y} \setminus Y_i) &= f(Y_i|Y_j : j \sim i) \\ &= \frac{\exp\left\{\sum_{k=0}^{p} \theta_k X_{k,i} Y_i + \sum_{j\sim i} \theta_{ij} Y_i(2Y_j - 1)\right\}}{1 + \exp\left\{\sum_{k=0}^{p} \theta_k X_{k,i} + \sum_{j\sim i} \theta_{ij}(2Y_j - 1)\right\}}, \end{aligned} \quad (1.2)$$

where $i = 1, \ldots, n$.

Autologistic regression models are suitable for relating a binary response variable to potential explanatory variables by a logistic regression, while accounting for spatial dependence by an autoregression. Moreover autologistic regression models can be used to estimate the probability of success at a given site and predict the outcome at an unsampled site. Hence autologistic regression models have been applied to many disciplines such as epidemiology, image analysis, and environmental studies (see, e.g., Besag, York, and Mollie 1991; Wu and Huffer 1997; Huffer and Wu 1998; Hoeting, Leecaster, and Bowden 2000). In particular, Gumpertz, Graham, and Ristaino (1997) gave an excellent account of autologistic models with regression and analyzed the spatial pattern of a Phytophthora epidemic in bell pepper. However, the aforementioned autologistic regression model is suitable for binary data on a spatial lattice at a given time point. Oftentimes observations are taken repeatedly over time and binary data are available on the same spatial lattice at multiple time points. That is, for a given location $\mathbf{s}_i$ and a given time point t, the response variable is $Y_{i,t} \equiv Y(\mathbf{s}_i, t)$, where $i = 1, \ldots, n$ and $t = 1, 2, \ldots$.

In this article, we propose a general spatial-temporal autologistic regression model as an extension of the (atemporal) autologistic regression model. The spatial-temporal autologistic regression model captures both spatial dependence and temporal dependence simultaneously by a space-time Markov random field, in addition to a logistic regression on potential explanatory variables. As we shall demonstrate in a data example, our generalized model has good potential in capturing correlation across space and over time. There is also evidence that it can give credible prediction of future responses. For statistical inference, we use maximum pseudo-likelihood, which is computationally efficient for parameter estimation, and develop a Markov chain Monte Carlo (MCMC) algorithm for predicting future responses.

The formulation of the model bears similarity to Besag (1972) and Preisler (1993). In Besag (1972), a spatial-temporal autologistic model was proposed with the assumption of stationarity and hence might not be suitable for incorporating regression terms. On the other hand, Preisler (1993) considered spatial-temporal autologistic regression, but assumed independence among different time points. Our contribution here is to improve upon Preisler's model so that the temporal dependence is accounted for in our proposed spatial-temporal autologistic regression model. Furthermore, our method can be extended to more general Markov random fields with pairwise spatial and temporal dependence.

The remainder of the article is organized as follows. Section 2 proposes the spatial-temporal autologistic regression model, estimates model parameters by maximum pseudo-likelihood, and uses an MCMC algorithm for prediction. Section 3 applies the spatial-temporal autologistic regression model to study the outbreaks of southern pine beetle in North Carolina. Section 4 discusses further model generalization.

2. SPATIAL-TEMPORAL AUTOLOGISTIC REGRESSION MODEL

Consider a binary spatial-temporal process $\{Y_{i,t} : i = 1, \ldots, n, t \in \mathbb{Z}\}$, where $Y_{i,t} \equiv Y(\mathbf{s}_i, t) = 0$ or 1 corresponds to the ith site $\mathbf{s}_i$ and time point t with $i = 1, \ldots, n$ and $t \in \mathbb{Z}$. For a given time point t, let $\mathbf{Y}_t \equiv (Y_{1,t}, \ldots, Y_{n,t})'$ denote the binary responses on the lattice $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$.

We propose to model the joint distribution of $\{\mathbf{Y}_t : t \in \mathbb{Z}\}$ by specifying a family of conditional distributions:

$$
\begin{aligned}
f(\mathbf{Y}_{t_1}, &\ldots, \mathbf{Y}_{t_2} | \{\mathbf{Y}_t : t \in \mathbb{Z} \setminus \{t_1, \ldots, t_2\}\}) \\
\propto\ & \exp \left\{ \sum_{t'=t_1}^{t_2} \left(\sum_{i=1}^{n} \sum_{k=0}^{p} \theta_k X_{k,i,t'} Y_{i,t'} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j \sim i} \theta_{p+1} [Y_{i,t'} Y_{j,t'} \right.\right. \\
& + (1 - Y_{i,t'})(1 - Y_{j,t'})] \\
& \left.\left. + \sum_{i=1}^{n} \theta_{p+2} [Y_{i,t'}(Y_{i,t'-1} + Y_{i,t'+1}) + (1 - Y_{i,t'})(2 - Y_{i,t'-1} - Y_{i,t'+1})] \right) \right\},
\end{aligned}
$$
(2.1)

for all $t_1, t_2 \in \mathbb{Z}$ such that $t_1 \leq t_2$, where $X_{k,i,t} \equiv X_k(\mathbf{s}_i, t)$ denotes the kth explanatory variable at site $\mathbf{s}_i$ and time point t, and θ_k is the logistic regression coefficient corresponding to $X_k(\cdot); k = 0, \ldots, p$. Further, θ_{p+1} is the spatial autoregression coefficient and θ_{p+2} is the temporal autoregression coefficient. Note that the specification is consistent for all $t_1 \leq t_2$, and the joint distribution of $\{\mathbf{Y}_t : t \in \mathbb{Z}\}$ can be shown to exist by Theorem 2.1.1 of Guyon (1995). In this article, we restrict our attention to space and time invariant logistic regression coefficients and autoregression coefficients.

Now for the ith site and the tth time point, define a neighborhood set

$$
N_{i,t} \equiv \{(j,t) : j \sim i\} \cup \{(i, t-1), (i, t+1)\}; \quad i = 1, \ldots, n, t \in \mathbb{Z}. \quad (2.2)
$$

From (2.1), it follows directly that the full conditional distribution of $Y_{i,t}$ is:

$$
\begin{aligned}
f\left(Y_{i,t} | \{\mathbf{Y}_t : t \in \mathbb{Z}\} \setminus Y_{i,t}\right) &= f\left(Y_{i,t} | Y_{j,t} : (j,t) \in N_{i,t}\right) \\
&= \frac{\exp\left\{ \sum_{k=0}^{p} \theta_k X_{k,i,t} Y_{i,t} + \sum_{j \sim i} \theta_{p+1} Y_{i,t}(2Y_{j,t} - 1) + \theta_{p+2} Y_{i,t}(2Y_{i,t-1} + 2Y_{i,t+1} - 2) \right\}}{1 + \exp\left\{ \sum_{k=0}^{p} \theta_k X_{k,i,t} + \sum_{j \sim i} \theta_{p+1}(2Y_{j,t} - 1) + \theta_{p+2}(2Y_{i,t-1} + 2Y_{i,t+1} - 2) \right\}},
\end{aligned}
$$
(2.3)

where $i = 1, \ldots, n, t \in \mathbb{Z}$.

Note that the difference between (1.2) and (2.3) is the temporal term. Hence the interpretation of the regression coefficients θ_k with $k = 0, \ldots, p$ and the spatial autoregression coefficient θ_{p+1} is similar to that of the usual (atemporal) autologistic model. In particular, θ_k represents the changes in the log conditional odds of an outbreak for a unit change in the corresponding explanatory variable X_k for $k = 0, \ldots, p$ and θ_{p+1} can be thought of as

a spatial dependence parameter (see, e.g., Cressie 1993, p. 424). The additional parameter θ_{p+2} is the temporal autoregression coefficient and can be thought of as a temporal dependence parameter, in a way similar to the spatial autoregression coefficient θ_{p+1}. When $\theta_{p+2} = 0$, there is no correlation over time, whereas when $\theta_{p+2} \neq 0$, there is correlation over time. A positive θ_{p+2} typically corresponds to a positive temporal correlation while a negative θ_{p+2} typically corresponds to a negative temporal correlation. The magnitude of θ_{p+2} is related to the mean difference between consecutive time points at the same site with same values ((0,0) or (1,1)) and those with opposite values ((0,1) or (1,0)).

The marginal logistic models used in Gumpertz et al. (2000) focused on the relationship between the explanatory variables and the probability of outbreaks while the spatial and temporal correlations were of secondary interests. In contrast, the idea of the spatial-temporal autologistic regression models here is to model the relationship between the explanatory variables and the probability of outbreaks while accounting for the spatial-temporal dependence simultaneously. If the primary interest of a study is in the regression terms, then both the marginal logistic regression models and the spatial-temporal autologistic regression models would be appropriate. If, in addition, it is of interest to understand the spatial-temporal dependence structure and make predictions at unsampled locations and into the future, then the spatial-temporal autologistic regression models would perhaps be more suitable than the marginal logistic regression models.

2.1   Parameter Estimation by Maximum Pseudo-likelihood

Corresponding to the model specified in (2.1), denote the model parameters by $\boldsymbol{\theta} \equiv (\theta_0, \theta_1, \dots, \theta_{p+2})'$. Suppose observations are obtained from T time points: $\mathbf{Y}_1, \dots, \mathbf{Y}_T$, where $\mathbf{Y}_t = (Y_{1,t}, \dots, Y_{n,t})'; t = 1, \dots, T$. To avoid dealing with the complex distributions of $\mathbf{Y}_1$ and $\mathbf{Y}_T$ at the end time points, we consider the following likelihood function of $\boldsymbol{\theta}$ based on the joint distribution of $\mathbf{Y}_2, \dots, \mathbf{Y}_{T-1}$ conditional on $\mathbf{Y}_1$ and $\mathbf{Y}_T$:

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}_2, \dots, \mathbf{Y}_{T-1} | \mathbf{Y}_1, \mathbf{Y}_T) \\
&= c(\boldsymbol{\theta})^{-1} \exp \left\{ \sum_{t=2}^{T-1} \left(\sum_{i=1}^{n} \sum_{k=0}^{p} \theta_k X_{k,i,t} Y_{i,t} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j \sim i} \theta_{p+1} [Y_{i,t} Y_{j,t} \right. \right. \\
&\quad + (1 - Y_{i,t})(1 - Y_{j,t})] \\
&\quad + \left. \left. \sum_{i=1}^{n} \theta_{p+2} [Y_{i,t}(Y_{i,t-1} + Y_{i,t+1}) + (1 - Y_{i,t})(2 - Y_{i,t-1} - Y_{i,t+1})] \right) \right\} . \quad (2.4)
\end{aligned}
$$

Here since the normalizing constant $c(\boldsymbol{\theta})$ does not have a closed form, direct maximization of the likelihood (2.4) would require approximation of $c(\boldsymbol{\theta})$ by, for example, a path sampling technique using Markov chain Monte Carlo (MCMC) (see, e.g., Gelman and Meng 1998). Because the MCMC requires intensive computations, we use "maximum pseudo-likelihood" for parameter estimation, which is easier to compute (Besag 1975). The pseudo-likelihood function, under our context, is the product of the full conditional distributions $f(Y_{i,t} | Y_{j,t} : (j, t) \in N_{i,t})$ as in (2.3); $i = 1, \dots, n, t = 2, \dots, T - 1$.

Maximization of the pseudo-likelihood function could be processed to obtain the maximum pseudo likelihood estimates (MPLE) $\hat{\boldsymbol{\theta}}$ by a standard logistic regression software routine such as `proc logistic` in SAS or `glm()` in S-Plus. Moreover, the maximum pseudo-likelihood estimates (MPLE) are generally consistent and asymptotically normal for Markov random fields (see, e.g., Guyon 1995). For autologistic models (with spatial dependence but without explanatory variables), the efficiency of MPLE depends on the values of the spatial autocorrelation coefficient and can at times be comparable to the efficiency of the maximum likelihood estimates (see Gumpertz et al. 1997, sec. 3).

However, the standard deviations of these estimates from the standard logistic regression are invalid and hence need to be assessed differently. We use a parametric bootstrap in a manner similar to Gumpertz et al. (1997). In particular, we generate M spatial-temporal binary datasets according to the autologistic model defined in (2.1), for which the model parameters are fixed at the MPLE $\hat{\boldsymbol{\theta}}$ from the original data. For the mth dataset, we compute the MPLE $\hat{\boldsymbol{\theta}}^{(m)}$, for $m = 1, \ldots, M$. The standard deviation of these MPLE's $\{\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(M)}\}$ can be used to estimate the standard deviation of $\hat{\boldsymbol{\theta}}$.

To generate a spatial-temporal binary dataset $\{\mathbf{Y}_2, \ldots, \mathbf{Y}_{T-1}\}$ given $\mathbf{Y}_1$ and $\mathbf{Y}_T$, we use (2.4). The normalizing constant $c(\boldsymbol{\theta})$ does not have a closed form and hence direct sampling of $\mathbf{Y}_2, \ldots, \mathbf{Y}_{T-1}$ from $f(\mathbf{Y}_2, \ldots, \mathbf{Y}_{T-1}|\mathbf{Y}_1, \mathbf{Y}_T)$ is not possible. Therefore we use Markov chain Monte Carlo, or more specifically, a Gibbs sampler. The main idea is to successively simulate individual $Y_{i,t}$ from the full conditional distribution as in (2.3) for $i = 1, \ldots, n, t = 2, \ldots, T-1$, and hence obtain a Markov chain that converges to the target distribution $f(\mathbf{Y}_2, \ldots, \mathbf{Y}_{T-1}|\mathbf{Y}_1, \mathbf{Y}_T)$. After burn-in, we take M samples from the Markov chain as the bootstrap samples. We use the log-likelihood values to determine the length of burn-in iterations as in Geweke (1992).

2.2 OPTIMAL PREDICTION BY MARKOV CHAIN MONTE CARLO

As in (2.4), for predicting $\mathbf{Y}_{T+1}, \ldots, \mathbf{Y}_{T^*}$, we consider the joint predictive distribution of $\mathbf{Y}_{T+1}, \ldots, \mathbf{Y}_{T^*+T^{**}-1}$ conditioned on the observation $\mathbf{Y}_T$ and a prespecified value $\mathbf{Y}_{T^*+T^{**}}$:

$$
\begin{aligned}
f(\mathbf{Y}_{T+1}, &\ldots, \mathbf{Y}_{T^*+T^{**}-1}|\mathbf{Y}_T, \mathbf{Y}_{T^*+T^{**}}) \\
\propto \quad &\exp \left\{ \sum_{t=T+1}^{T^*+T^{**}-1} \left(\sum_{i=1}^{n} \sum_{k=0}^{p} \theta_k X_{k,i,t} Y_{i,t} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j \sim i} \theta_{p+1}[Y_{i,t}Y_{j,t} \right. \right. \\
&+ (1 - Y_{i,t})(1 - Y_{j,t})] \\
&\left. \left. + \sum_{i=1}^{n} \theta_{p+2}[Y_{i,t}(Y_{i,t-1} + Y_{i,t+1}) + (1 - Y_{i,t})(2 - Y_{i,t-1} - Y_{i,t+1})] \right) \right\}, (2.5)
\end{aligned}
$$

where T^{**} extra time points are included to reduce the potential boundary effects at time point T^*, with $T+1 \leq T^* \leq T^*+T^{**}-1$. Again we use a Gibbs sampler to draw samples $\mathbf{Y}_{T+1}, \ldots, \mathbf{Y}_{T^*+T^{**}-1}$ from the predictive distribution $f(\mathbf{Y}_{T+1}, \ldots, \mathbf{Y}_{T^*+T^{**}-1}|\mathbf{Y}_T,$
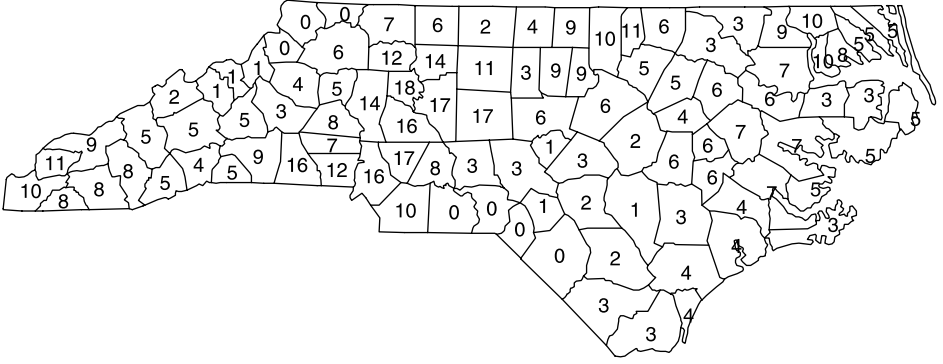
Figure 1. Total number of years a county experienced an outbreak of southern pine beetle in 1960–1996, for each of the 100 counties of North Carolina.

$\mathbf{Y}_{T^*+T^{**}}$). Similar to (2.3), it follows that the full conditional distribution of $Y_{i,t}$ is:

$$f(Y_{i,t}|\{\mathbf{Y}_T, \ldots, \mathbf{Y}_{T^*+T^{**}}\} \setminus Y_{i,t}) = f\left(Y_{i,t}|Y_{j,t} : (j,t) \in N_{i,t}\right)$$

$$= \frac{\exp\{\sum_{k=0}^{p} \theta_k X_{k,i,t} Y_{i,t} + \sum_{j\sim i} \theta_{p+1} Y_{i,t}(2Y_{j,t} - 1) + \theta_{p+2} Y_{i,t}(2Y_{i,t-1} + 2Y_{i,t+1} - 2)\}}{1 + \exp\{\sum_{k=0}^{p} \theta_k X_{k,i,t} + \sum_{j\sim i} \theta_{p+1}(2Y_{j,t} - 1) + \theta_{p+2}(2Y_{i,t-1} + 2Y_{i,t+1} - 2)\}},$$

where $i = 1, \ldots, n, T+1 \le t \le T^*+T^{**}-1$. Sampling from these individual conditional distributions is straightforward. Upon convergence, the samples of $\{Y_{i,t} : i = 1, \ldots, n, t = T+1, \ldots, T^*\}$ are used to generate the predicted binary responses.

3. EXAMPLE: OUTBREAKS OF SOUTHERN PINE BEETLE

In this section, we apply the spatial-temporal autologistic regression model to a study of the outbreaks of southern pine beetle (*Dendrotonus frontalis*) in North Carolina. Recall that Gumpertz et al. (2000) aggregated the binary data to a count at a given site and modeled the proportion of years each site experienced an outbreak. The temporal correlation was accounted for by an overdispersion in the working variance-covariance matrix using generalized estimating equations, while the spatial correlation was accounted for by the sample correlation between each pair of sites. However, as mentioned in Section 1, the analysis was performed in several steps. In this regard, the spatial-temporal autologistic regression model developed in Section 2 provides a systematic alternative to account for both spatial dependence and temporal dependence.

The data consist of the presence and absence of southern pine beetle in the 100 counties of North Carolina from 1960 to 1996. That is, $\{Y_{i,t} : i = 1, \ldots, 100, t = 1960, \ldots, 1996\}$, where $Y_{i,t} = 0$ for absence and $Y_{i,t} = 1$ for presence of an outbreak in the ith county and the tth year. We used the first 31 years (1960–1990) of data for model building and set aside
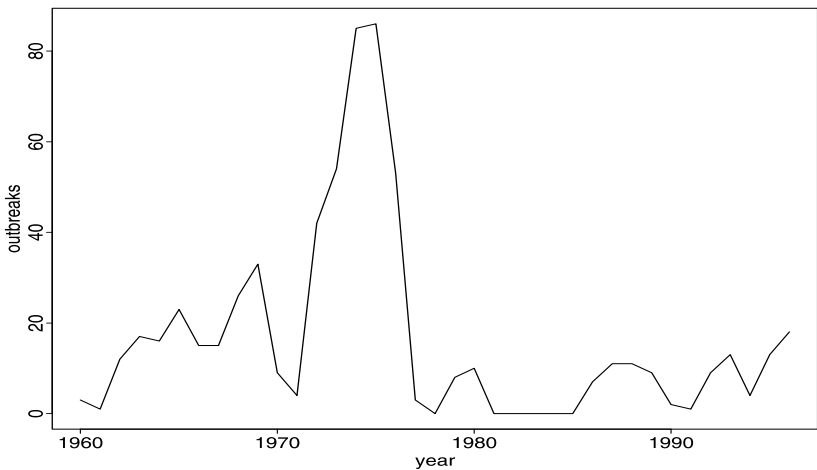
Figure 2. Total number of counties that experienced an outbreak of southern pine beetle in the state of North Carolina from 1960 to 1996.

the last 6 years (1991–1996) of data for model validation, as in Gumpertz et al. (2000). Two counties were considered to be neighbors if the corresponding county seats are within 30 miles of each other, similar to the neighborhood structure considered in Section 6.1 of Cressie (1993). Figure 1 plots the total number of years a county experienced an outbreak for each of the 100 counties. The spatial distribution demonstrates positive correlation among neighboring counties. Indeed, Moran's I index is .64 with a p value less than .001 and Geary's C index is .32 with a p value less than .001, both indicating some evidence of positive spatial correlation. Figure 2 is a time-series plot of the total number of counties that experienced an outbreak in a year, for each of the years from 1960–1996. The epidemic seems to have peaked in the mid-1970s and there is evidence of positive correlation over time.

Among the possible explanatory variables, we focused on the 11 most important explanatory variables identified by Gumpertz et al. (2000): elevation (in m), longitude, saw volume (in m³/ha), hydric proportion, xeric proportion, size of national forest (in 1,000 ha), average daily maximum temperature in the fall (in °C), average precipitation in the fall (in cm), average daily maximum temperature in the winter (in °C), average daily maximum temperature in the summer (in °C), and average precipitation in the summer (in cm). These variables were recorded at the county level and some of the variables were transformed to either a log or square-root scale. Two interactions, one between the saw volume and the average daily maximum winter temperature and the other between the saw volume and the average daily maximum summer temperature, were created as in Gumpertz et al. (2000). Along with the spatial component and the temporal component, there are a total of 15 variables in the autologistic model (Table 1).

The MPLE $\hat{\theta}$ was obtained by maximizing the product of the full conditional distribu-

Table 1. Maximum Pseudo-Likelihood Estimates of the Coefficients for the Final Model Selected by Backward Elimination

Variable	Estimate	Bootstrap bias	Bootstrap SD
X_0 Intercept	−28.492	−.408	8.093
X_1 Ln[elevation (m)]	—	—	—
X_2 Longitude	—	—	—
X_3 $\sqrt{\text{saw volume (m}^3\text{/ha)}}$	1.318	.001	.772
X_4 $\sqrt{\text{hydric proportion}}$	−.068	−.005	.056
X_5 $\sqrt{\text{xeric proportion}}$	.040	.003	.033
X_6 $\sqrt{\text{national forest (thousand ha)}}$	—	—	—
X_7 Mean daily maximum fall temp (C)	−.249	−.003	.153
X_8 Mean fall precipitation (cm)	.666	.030	.250
X_9 Mean daily maximum winter temp (C)	—	—	—
X_{10} Mean daily maximum summer temp (C)	.515	.006	.193
X_{11} Mean summer precipitation (cm)	—	—	—
$X_{12} = X_3 \times X_9$	—	—	—
$X_{13} = X_3 \times X_{10}$	−.015	.000	.009
Spatial effect	.807	.012	.088
Temporal effect	.810	−.013	.121

tions:

$$f\left(Y_{i,t}|Y_{j,t} : (j,t) \in N_{i,t}\right)$$
$$= \frac{\exp\left\{\sum_{k=0}^{13} \theta_k X_{k,i,t} Y_{i,t} + \theta_{14} \sum_{j\sim i} Y_{i,t}(2Y_{j,t} - 1) + \theta_{15} Y_{i,t}(2Y_{i,t-1} + 2Y_{i,t+1} - 2)\right\}}{1 + \exp\left\{\sum_{k=0}^{13} \theta_k X_{k,i,t} + \theta_{14} \sum_{j\sim i}(2Y_{j,t} - 1) + \theta_{15}(2Y_{i,t-1} + 2Y_{i,t+1} - 2)\right\}},$$
(3.1)

for $i = 1, \ldots, 100$ and $t = 1961, \ldots, 1990$. Evaluated at the MPLE $\hat{\theta}$, a Gibbs sampler was implemented according to (3.1) and after burn-in, a bootstrap sample of size $M = 5,000$ was generated. From these 5,000 samples, the standard deviations of the MPLE $\hat{\theta}$ were estimated. The MPLEs and their corresponding standard deviations are reported in Table 2.

Because not all the parameter estimates $\{\hat{\theta}_k : k = 0, \ldots, 15\}$ were significantly different from zero, we set out to determine a suitable reduced model. We started with the full model (3.1) and performed backward elimination based on a t-ratio of an estimate $\hat{\theta}_k$ and its standard deviation. At each step, we eliminated the variable that had the least t-ratio and then fit the reduced model to the data using maximum pseudo-likelihood, as we did with the full model. We used a unit t-ratio as our cut-off, which has been reported effective for model selection in linear regression (see, e.g., Chatterjee, Hadi, and Price 2000, sec. 11.9). The elimination procedure was stopped when all the coefficients had t-ratios above 1. The steps in the backward elimination are shown in Table 2. In particular, the variables were eliminated in the following order: the mean summer precipitation (X_{11}), the size of national forest (X_6), the elevation (X_1), the mean daily maximum winter temperature (X_9), interaction between the saw volume and the mean daily maximum winter temperature (X_{12}), and finally the longitude (X_2). The final reduced model has seven explanatory variables, namely the saw volume (X_3), the hydric proportion (X_4), the xeric proportion (X_5), the mean daily maximum fall temperature (X_7), the mean fall precipitation (X_8), the mean daily

Table 2. Individual Steps in a Backward Elimination. Reported are the maximum pseudo-likelihood estimates of the coefficients and the standard deviations (in parentheses) estimated by a bootstrap method.

Variable	Full model		Step 1		Step 2		Step 3	
X_0	−20.728	(16.904)	−21.515	(15.636)	−21.837	(15.102)	−23.848	(14.255)
X_1	.049	(.202)	.052	(.194)	.047	(.189)	—	—
X_2	.162	(.175)	.159	(.178)	.145	(.149)	.116	(.116)
X_3	1.739	(1.588)	1.829	(1.255)	1.776	(1.158)	1.725	(1.025)
X_4	−.113	(.069)	−.115	(.069)	−.115	(.067)	−.114	(.068)
X_5	.080	(.044)	.079	(.049)	.080	(.048)	.079	(.049)
X_6	.015	(.055)	.012	(.052)	—		—	
X_7	−.373	(.480)	−.358	(.459)	−.322	(.403)	−.346	(.384)
X_8	.720	(.355)	.708	(.324)	.703	(.298)	.704	(.292)
X_9	−.085	(.302)	−.102	(.263)	−.110	(.251)	−.113	(.246)
X_{10}	.726	(.425)	.729	(.413)	.695	(.361)	.717	(.354)
X_{11}	−.037	(.263)	—	—	—	—	—	—
X_{12}	.017	(.021)	.017	(.020)	.017	(.019)	.017	(.018)
X_{13}	−.031	(.025)	−.032	(.022)	−.031	(.022)	−.030	(.020)
X_{14}	.812	(.104)	.812	(.101)	.812	(.100)	.811	(.096)
X_{15}	.811	(.143)	.810	(.135)	.811	(.128)	.810	(.122)

Variable	Step 4		Step 5		Step 6	
X_0	−22.862	(12.805)	−22.669	(12.712)	−28.492	(8.093)
X_1	—		—		—	
X_2	.114	(.118)	.076	(.098)	—	
X_3	1.657	(.944)	1.308	(.932)	1.318	(.772)
X_4	−.110	(.066)	−.093	(.063)	−.068	(.056)
X_5	.073	(.048)	.063	(.040)	.040	(.033)
X_6	—		—		—	
X_7	−.459	(.310)	−.273	(.157)	−.249	(.153)
X_8	.701	(.270)	.696	(.244)	.666	(.250)
X_9	—		—		—	
X_{10}	.729	(.350)	.536	(.207)	.515	(.193)
X_{11}	—		—		—	
X_{12}	.011	(.014)	—		—	
X_{13}	−.026	(.018)	−.015	(.011)	−.015	(.009)
X_{14}	.811	(.090)	.809	(.095)	.807	(.088)
X_{15}	.813	(.117)	.812	(.130)	.810	(.121)

maximum summer temperature (X_{10}), and interaction between the saw volume and the mean daily maximum summer temperature (X_{13}). Interestingly both the spatial component and the temporal component were retained. In fact, both components were the most significant variables throughout the model-selection steps. The log-odds of outbreak in the fitted final model is:

$$\log\left(\frac{\Pr(Y_{i,t} = 1 | Y_{j,t} : (j,t) \in N_{i,t})}{\Pr(Y_{i,t} = 0 | Y_{j,t} : (j,t) \in N_{i,t})}\right)$$

$$= -28.492 + 1.318 \times \sqrt{\text{saw volume}} - .068 \times \sqrt{\text{hydric proportion}}$$

$$+ .040 \times \sqrt{\text{xeric proportion}} - .249 \times \text{fall temp} + .666 \times \text{fall precip}$$

$$+ .515 \times \text{summer temp} - .015 \times \sqrt{\text{saw volume}} * \text{summer temp}$$

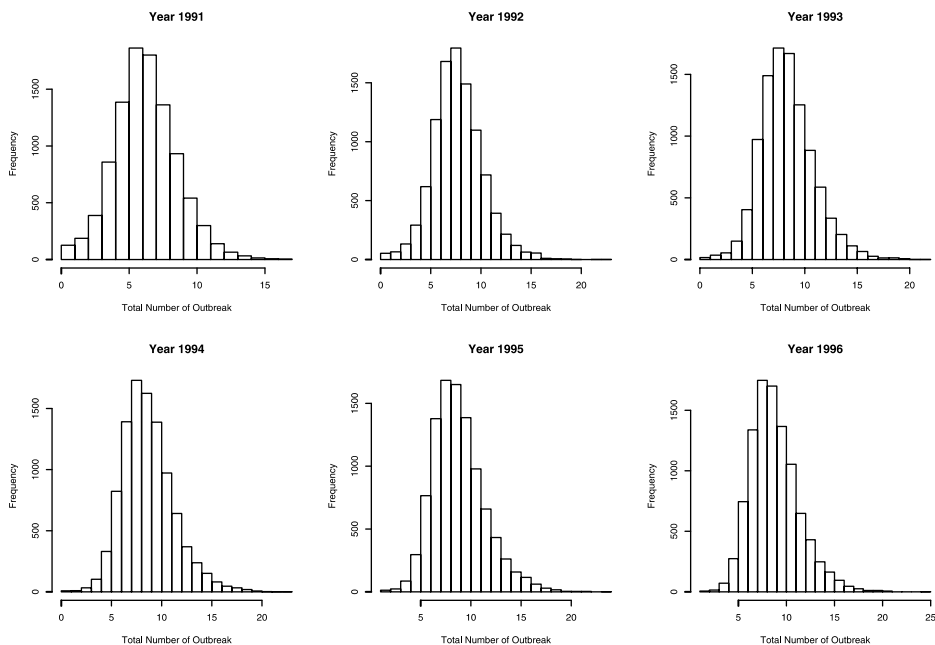$$+ .807 \times \sum_{j \sim i} (2Y_{j,t} - 1) + .810 \times (2Y_{i,t-1} + 2Y_{i,t+1} - 2). \qquad (3.2)$$

Figure 3. Histograms of the predicted total number of counties that experienced an outbreak of southern pine beetle in the state of North Carolina from 1991 to 1996.

The MPLEs and their corresponding standard deviations for the final model are reported in Table 1. As a byproduct, the empirical bias of the MPLEs based on the bootstrap samples was computed and was found to be negligible.

Given the fitted parameter values of the final model, we then used a Gibbs sampler to obtain the prediction of outbreaks from 1991 to 1996. In particular, we considered the joint predictive distribution of $\mathbf{Y}_{1991}, \ldots, \mathbf{Y}_{2000}$, conditioned on $\mathbf{Y}_{1990}$ and $\mathbf{Y}_{2001}$ as in (2.5). That is, with $T = 1990, T^* = 1996$, we chose $T^{**} = 5$ to reduce the effect of the boundary time point on the prediction. The Gibbs sampler was based on the full conditional distributions $f(Y_{i,t}|\{\mathbf{Y}_{1990}, \ldots, \mathbf{Y}_{2001}\}\backslash Y_{i,t})$ as described in (3.2), for $t = 1991, \ldots, 2000$. Upon convergence of the Gibbs sampler, $M = 10{,}000$ sets of $\{Y_{i,t} : i = 1, \ldots, n, t = 1991, \ldots, 1996\}$ were used as the predicted binary responses. Here we generated $Y_{i,2001}$ for the ith county at the end time point as an independent trial with an outbreak probability of $(\sum_{t=1960}^{1990} Y_{i,t})/31$ (i.e., the average outbreak rate of this county in 1960–1990), where $i = 1, \ldots, 100$. Using a small simulation study, we assessed the effect of $\mathbf{Y}_{T^*+T^{**}} = \mathbf{Y}_{2001}$ and found that in this case the values of $\mathbf{Y}_{2001}$ had little effect on the prediction of $\mathbf{Y}_{1991}, \ldots, \mathbf{Y}_{1996}$. Further, the explanatory variables in 1991–1996 were used for prediction, but not the actual observed outbreaks, as the terms $\mathbf{Y}_{1991}, \ldots, \mathbf{Y}_{1996}$ that appear in the full conditional distributions are simulated values in the Gibbs sampler iterations.

Figure 3 shows the histograms of the $M = 10{,}000$ predicted total number of outbreaks in North Carolina for each year in 1991–1996. Figure 4 shows the average predicted total number of outbreaks in 1991–1996 for each county in North Carolina. Using the actual
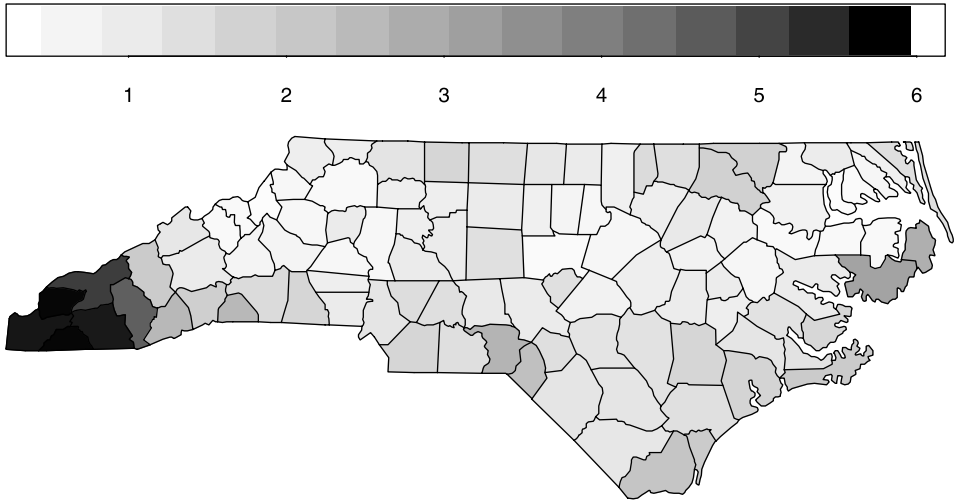
Figure 4. Mean predicted total number of years a county experienced an outbreak of southern pine beetle in 1991–1996, for each of the 100 counties of North Carolina.

data from 1991–1996, we computed a prediction error rate as the proportion of counties for which an outbreak was predicted differently from the actual observation, for each set of $\{Y_{i,t} : i = 1, \ldots, n\}$, where $t = 1991, \ldots, 1996$. We used two ways to compute the predicted value in the prediction error rate based on the $M = 10{,}000$ Gibbs samples: one is the mode of the predictive distribution and the other is the mean of the predictive distribution rounded to an integer value, for each $i = 1, \ldots, n$. When the mode of the Gibbs sample was used, the prediction error rates were .05, .15, .19, .06, .17, and .24 for 1991–1996; whereas when the mean of the Gibbs sample was used, the prediction error rates were .08, .17, .22, .09, .20, and .27. Our model gave reasonably good predictions. This, and the fact that the spatial and temporal components played an important role in the final reduced model gave justification for the need of the spatial-temporal autologistic regression model developed in this article.

4. DISCUSSION

In this article, we have developed an autologistic regression model for binary data which allows for a logistic regression while accounting for both spatial dependence and temporal dependence. We have used maximum pseudo-likelihood for parameter estimation and a parametric bootstrap for the corresponding standard deviations. Further we have proposed a Gibbs sampler to predict the responses at future time points. The methodology has been applied to successfully characterize and predict the outbreaks of southern pine beetle in North Carolina, based on 31 years of data. It is worth mentioning that the explanatory variables in our dataset are time-invariant, including the weather information. We believe that the fit of the model and the prediction of outbreaks could be improved further, should

those time-variant explanatory variables become available.

Note that the spatial model (1.1) can be reparameterized to:

$$f(\mathbf{Y}) \propto \exp\left\{ \sum_{i=1}^{n} \sum_{k=0}^{p} \theta_k^* X_{k,i} Y_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j\sim i} \theta_{ij}^* Y_i Y_j \right\}.$$

When $X_{0,i} \equiv 1$ for all i and thus θ_0^* corresponds to an intercept, there is a 1:1 correspondence between the parameters $\{\theta_k, \theta_{ij}\}$ and $\{\theta_k^*, \theta_{ij}^*\}$. Similar reparameterization can be applied to the spatial-temporal model (2.1).

Our approach can be extended to Gibbs fields to form more general spatial-temporal automodels with pairwise spatial-temporal dependence. More specifically, the joint distribution of $\{\mathbf{Y}_t : t \in \mathbb{Z}\}$ can be specified via a family of conditional distributions in a Gibbsian form. Examples of Gibbs fields include autobinomial, negative-binomial, Poisson, and Gaussian models. With proper parameterization, statistical inference can be carried out in a similar manner as in Section 2. It may also be of interest to extend the model to have time-varying coefficients and space-time interaction terms, which we leave for future investigation.

ACKNOWLEDGMENTS

[Received May 2003. Revised November 2004.]

REFERENCES

Bailey, R. (1995), "Description of the Ecoregions of the United States," *Misc. Publication 1391*, Washington, DC: USDA Forest Service, p. 108.

Besag, J. (1972), "Nearest-Neighbour Systems and the Auto-logistic Model for Binary Data," *Journal of the Royal Statistical Society*, Ser. B, 34, 75–83.

——— (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society*, Ser. B, 36, 192–225.

——— (1975), "Statistical Analysis of Non-lattice Data," *The Statistician*, 24, 179–195.

Besag, J., York, J., and Mollie, A. (1991), "Bayesian Image Restoration With Two Applications in Spatial Statistics," *Annals of the Institute of Mathematical Statistics*, 43, 1–59.

Chatterjee, S., Hadi, A. S., and Price, B. (2000), *Regression Analysis by Example* (3rd ed.), New York: Wiley.

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

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

Gelman, A., and Meng, X.-L. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press.

Greig, D. M., Porteous, B. T., and Seheult, A. H. (1989), "Exact Maximum a posteriori Estimation for Binary Images," *Journal of the Royal Statistical Society*, Ser. B, 51, 271–279.

Gumpertz, M. L., Graham, J. M., and Ristaino, J. B. (1997), "Autologistic Model of Spatial Pattern of Phytophthora Epidemic in Bell Pepper: Effects of Soil Variables on Disease Presence," *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 131–156.

Gumpertz, M. L., Wu, C.-T., and Pye, J. M. (2000), "Logistic Regression for Southern Pine Beetle Outbreaks with Spatial and Temporal Autocorrelation," *Forest Science*, 46, 95–107.

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

Hoeting, J. A., Leecaster, M., and Bowden, D. (2000), "An Improved Model for Spatially Correlated Binary Responses," *Journal of Agricultural, Biological, and Environmental Statistics*, 5, 102–114.

Huffer, F. W., and Wu, H. (1998), "Markov Chain Monte Carlo for Autologistic Regression Models With Application to the Distribution of Plant Species," *Biometrics*, 54, 509–535.

Mawby, W., and Gold, H. (1984), "A Reference Curve and Space-Time Series Analysis of the Regional Population Dynamics of the Southern Pine Beetle," *Research on Population Ecology*, 26, 261–274.

Preisler, H. K. (1993), "Modelling Spatial Patterns of Trees Attacked by Bark-Beetles," *Applied Statistics*, 42, 501–514.

Pye, J. M. (1993), "Regional Dynamics of Southern Pine Beetle Populations," *Spatial Analysis and Forest Pest Management*, eds. A. M. Liebhold and H. R. Barrett, USDA Forest Service General Technical Report, NE-175.

Turchin, P., Lorio, P., Taylor, A., and Billings, R. (1991), "Why do Populations of Southern Pine Beetles (Coleoptera: Scolytidae) Fluctuate?" *Environmental Entomology*, 20, 401–409.

Wu, H., and Huffer, F. W. (1997), "Modeling the Distribution of Plant Species using the Autologistic Regression Model," *Environmental and Ecological Statistics*, 4, 49–64.