Modeling Spatial-Temporal Binary Data Using Markov Random Fields

Jun Zhu
Department of Statistics
University of Wisconsin – Madison
Madison, WI 53706, USA
jzhu@stat.wisc.edu

Hsin-Cheng Huang
Institute of Statistical Science
Academia Sinica
Taipei 115, Taiwan
hchuang@stat.sinica.edu.tw

Chi-tsung Wu
Department of Statistics
Feng Chia University
Taichung, Taiwan
cwu@fcu.edu.tw

Abstract

An autologistic regression model consists of a linear regression of a response variable on explanatory variables and an auto-regression 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 model for spatial-temporal binary data. The model we propose is a Markov chain across time, where the transition probability is autologistic on the given lattice. 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 Beetles in North Carolina. We also discuss the generality of our approach for modeling other types of spatial-temporal lattice data.

Keywords and Phrases: Autologistic model, Gibbs sampler, Markov chain Monte Carlo, maximum pseudo-likelihood, spatial-temporal model.

1 Introduction

The Southern Pine Beetles have caused severe damages to pine forests in the southern states of the United States and hence are 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 Beetles 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 6-7 years for the outbreaks in the southern United States; Turchin, Lorio, Taylor, and Billings (1991) found temporal autocorrelation at a lag of 1–2 years for some populations in eastern Texas.

To our knowledge, Gumpertz, Wu, and Pye (2000) were the first to develop a statistical model which accounts for spatial and temporal autocorrelations, in addition to the potential explanatory variables. They studied the outbreaks of Southern Pine Beetles 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), the spatial dependence was accounted for specifically by an autologistic regression model, but the temporal aspect was handled implicitly by considering the total number of outbreaks over time at a given site. Statistical models were constructed by first assuming independence, then incorporating temporal dependence, and finally accounting for spatial dependence. 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 of the final inference. The primary purpose of this article is to develop a spatial-temporal autologistic model that would systematically account for spatial dependence and temporal dependence, *simultaneously*.

Figure 1–2 here

Our approach will be to add a temporal component to the existing atemporal autologistic model. Recall that autologistic models (Besag 1972, 1974) account for spatial dependence among binary variables on a regular or irregular lattice. With an addition of a linear regression, autologistic models can be used to model relationships between the binary response variable and potential

explanatory variables, while incorporating spatial correlation (see, e.g., Section 6.5.1, Cressie 1993). Consider representative sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ on a spatial lattice. For a given neighborhood structure, let $N_i \equiv \{j : \boldsymbol{s}_j \text{ is a neighbor of } \boldsymbol{s}_i\}$. 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 (or rook's case), diagonal (or bishop's case), and second order (or queen's case). Let $Y_1, \ldots, Y_n$ denote binary responses on the lattice, where $Y_i \equiv Y(\boldsymbol{s}_i) = 0$ or 1. The joint distribution of $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_n)'$ for an autologistic model can be formulated in a similar way as Greig, Porteous, and Seheult (1989):

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

where $X_{k,i} \equiv X_k(\boldsymbol{s}_i)$ denotes the k-th explanatory variable at site $\boldsymbol{s}_i$, θ_k denotes the k-th linear regression coefficient corresponding to $X_k(\cdot)$, with $k = 1, \ldots, p$. Further θ_{ij} denotes the autoregression coefficient between the i-th site and the j-th site, such that $\theta_{ij} = \theta_{ji}$ and $\theta_{ij} > 0$ if $j \sim i$. It follows from (1) that the conditional distributions are:

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

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

Autologistic models are suitable for relating a binary response variable to potential explanatory variables by a linear regression, while accounting for spatial dependence by an auto-regression. Moreover autologistic models can be used to estimate the probability of success at a given site and predict the outcome at an unsampled site. Hence autologistic 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). In particular, Gumpertz, Graham, and Ristaino (1997) gave an excellent account of autologistic models with regression and analyzed the spatial pattern of the Phytophthora epidemic in bell pepper. However, the aforementioned autologistic 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 $\boldsymbol{s}_i$ and a given time point t, the response variable is $Y_{i,t} \equiv Y(\boldsymbol{s}_i, t)$, where $i = 1, \ldots, n$ and $t = 1, 2, \ldots$.

In this article, we propose a general spatial-temporal autologistic model as an extension of the (atemporal) autologistic model. The spatial-temporal autologistic model captures spatial dependence by a Markov random field and captures temporal dependence explicitly by a Markov chain. The formulation bears similarity as Section 4 of Besag (1972), but is quite different. Besag (1972) considered an autologistic model and chose a transition probability for the Markov chain in such a way that a stationary distribution exists and is auto-logistic. It is not clear how the approach could be applied to autologistic model with regression, because explanatory variables can be time variant resulting in non-stationarity. Instead we model directly the transition probability by an autologistic model with regression. As we shall demonstrate in a data example, our generalized model has good potential in capturing correlations across space and over time. There is also evidence that it can give credible forecasting of future responses. For statistical inference, we use maximum pseudo-likelihood, which is computationally efficient, for parameter estimation (e.g., Gumpertz et al. 1997), and develop a Markov chain Monte Carlo (MCMC) algorithm for predicting future responses. The formulation of the model, to our knowledge, is novel. Furthermore, the method can be extended to more general Markov random fields with pairwise spatial dependence.

The remaining of the article is organized as follows. In Section 2, we propose the spatial-temporal autologistic model, estimate model parameters by maximum pseudo-likelihood, and use an MCMC algorithm for prediction. In Section 3, the spatial-temporal autologistic model is applied to study the outbreaks of Southern Pine Beetles in North Carolina. Discussion is given about further model generalization in Section 4.

2 Spatial-Temporal Autologistic Model

Consider a binary spatial-temporal process $\{Y_{i,t} : i = 1, \ldots, n, t = 1, 2, \ldots\}$, where $Y_{i,t} \equiv Y(\boldsymbol{s}_i, t) =$ 0 or 1 corresponds to the i-th site $\boldsymbol{s}_i$ and time point t with $i = 1, \ldots, n$ and $t = 1, 2, \ldots$. For a given time point t, let $\boldsymbol{Y}_t \equiv (Y_{1,t}, \ldots, Y_{n,t})'$ denote the binary responses on the lattice $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$. We model $\{\boldsymbol{Y}_t : t = 1, 2, \ldots\}$ by an n-dimensional vector Markov chain with the following transition probability:

$$f(\boldsymbol{Y}_t | \boldsymbol{Y}_{t-1}) = q(\boldsymbol{Y}_t | \boldsymbol{Y}_{t-1})/c. \tag{3}$$

Here c is a normalizing constant and

$$q(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1}) \equiv \exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{p}\theta_k X_{k,i,t}Y_{i,t} + \sum_{j\sim i}\theta_{p+1}[Y_{i,t}Y_{j,t} + (1-Y_{i,t})(1-Y_{j,t})] + \sum_{i=1}^{n}\theta_{p+2}[Y_{i,t}Y_{i,t-1} + (1-Y_{i,t})(1-Y_{i,t-1})]\right\}, \quad (4)$$

where $X_{k,i,t} \equiv X_k(\boldsymbol{s}_i, t)$ denotes the k-th explanatory variable at site $\boldsymbol{s}_i$ and time point t, θ_k is the linear regression coefficient corresponding to $X_k(\cdot)$; $k = 1, \ldots, p$. Further, θ_{p+1} is the spatial autoregression coefficient and θ_{p+2} is the temporal autoregression coefficient. In this article, we restrict our attention to space and time invariant regression coefficients and autoregression coefficients. We discuss possible relaxation of this assumption in Section 4.

For a given starting time point t_0, it follows from (3) and (4) that the joint distribution of $\boldsymbol{Y}_{t_0+1}, \ldots, \boldsymbol{Y}_t$ conditional on $\boldsymbol{Y}_{t_0}$ is,

$$f(\boldsymbol{Y}_{t_0+1}, \ldots, \boldsymbol{Y}_t|\boldsymbol{Y}_{t_0}) = \prod_{t'=t_0+1}^{t} f(\boldsymbol{Y}_{t'}|\boldsymbol{Y}_{t'-1}) = \frac{1}{c^{t-t_0}}\prod_{t'=t_0+1}^{t} q(\boldsymbol{Y}_{t'}|\boldsymbol{Y}_{t'-1})$$
$$= \frac{1}{c^{t-t_0}}\exp\left\{\sum_{t'=t_0+1}^{t}\left(\sum_{i=1}^{n}\sum_{k=1}^{p}\theta_k X_{k,i,t'}Y_{i,t'} + \sum_{j\sim i}\theta_{p+1}[Y_{i,t'}Y_{j,t'} + (1-Y_{i,t'})(1-Y_{j,t'})] + \sum_{i=1}^{n}\theta_{p+2}[Y_{i,t'}Y_{i,t'-1} + (1-Y_{i,t'})(1-Y_{i,t'-1})]\right)\right\}. \quad (5)$$

Now for the i-th site and the t-th time point, define a neighborhood set $N_{i,t} \equiv \{(j,t) : j \sim i\} \cup \{(i, t-1)\}$. From (5), our derivation (presented in the Appendix) shows that the full conditional distribution of $Y_{i,t}$ is:

$$f\left(Y_{i,t}|\{\boldsymbol{Y}_{t_0}, \ldots, \boldsymbol{Y}_t\} \setminus Y_{i,t}\right) = f\left(Y_{i,t}|Y_{j,t} : (j,t) \in N_{i,t}\right)$$
$$= \frac{\exp\left\{\sum_{k=1}^{p}\theta_k X_{k,i,t}Y_{i,t} + \sum_{j\sim i}\theta_{p+1}Y_{i,t}(2Y_{j,t}-1) + \theta_{p+2}Y_{i,t}(2Y_{i,t-1}-1)\right\}}{1 + \exp\left\{\sum_{k=1}^{p}\theta_k X_{k,i,t} + \sum_{j\sim i}\theta_{p+1}(2Y_{j,t}-1) + \theta_{p+2}(2Y_{i,t-1}-1)\right\}}, \quad (6)$$

where $i = 1, \ldots, n$ and $t = t_0 + 1, t_0 + 2, \ldots$.

Note that the difference between (1) and (4) (and between (2) and (6)) is the temporal term. Hence the interpretation of the regression coefficients θ_k with $k = 1, \ldots, p$ and the spatial autoregression coefficient θ_{p+1} is the same as that of the usual (atemporal) autologistic model. The additional parameter θ_{p+2} is the temporal autoregression coefficient. When $\theta_{p+2} = 0$, there is no correlation over time and the Markov chain of $\{\boldsymbol{Y}_t\}$ reduces to a sequence of independent random

vectors, each representing the usual autologistic lattice data at a given time point. Consequently $f(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1})$ in (3) reduces to $f(\boldsymbol{Y}_t)$ in (1). On the other hand, when $\theta_{p+2} \neq 0$, there is correlation over time. The magnitude of θ_{p+2} is related to the mean difference between two 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)).

Based on (6), the spatial-temporal autologistic model is Markovian in the sense that the response variable $Y_{i,t}$ at site $\boldsymbol{s}_i$ and time point t, conditional on all the other response variables at time points t and earlier, depends only on the response variables at the neighboring sites $\{Y_{j,t} : j \sim i\}$ and the response variable at the same site but from the previous time point $Y_{i,t-1}$. The joint distribution in (5), however, is assumed to be conditional on the response variables at the initial time point t_0, because it is non-trivial to specify the initial distribution $f(\boldsymbol{Y}_{t_0})$. We note that it is the transition probability $f(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1})$ that follows an autologistic model, not the marginal probability $f(\boldsymbol{Y}_t)$. Hence it might not suitable to let $f(\boldsymbol{Y}_{t_0})$ follow an autologistic model.

2.1 Parameter Estimation by Maximum Pseudo-likelihood

Corresponding to the model specified in (5), denote the model parameters by $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_{p+2})'$. Suppose observations are obtained from T time points: $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$, where $\boldsymbol{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})'$; $t = 1, \ldots, T$. From (5), the likelihood function of $\boldsymbol{\theta}$ is:

$$\begin{aligned} & \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T|\boldsymbol{Y}_1) = \prod_{t=2}^{T} f(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1}; \boldsymbol{\theta}) \\ = \quad & \frac{1}{(c(\boldsymbol{\theta}))^{T-1}} \exp\Bigg\{ \sum_{t=2}^{T} \Bigg(\sum_{i=1}^{n} \sum_{k=1}^{p} \theta_k X_{k,i,t} Y_{i,t} + \sum_{j \sim i} \theta_{p+1} [Y_{i,t} Y_{j,t} + (1 - Y_{i,t})(1 - Y_{j,t})] \\ & \qquad + \sum_{i=1}^{n} \theta_{p+2} [Y_{i,t} Y_{i,t-1} + (1 - Y_{i,t})(1 - Y_{i,t-1})] \Bigg) \Bigg\}. \end{aligned} \tag{7}$$

Since the normalizing constant $c(\boldsymbol{\theta})$ does not have a closed form, direct maximization of the likelihood (7) 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 estimates, which is easier to compute (Besag 1975). The pseudo-likelihood function, under our context, is the product of the conditional distributions in (6):

$$\mathcal{L}^*(\boldsymbol{\theta}; \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T|\boldsymbol{Y}_1) = \prod_{t=2}^{T} \prod_{i=1}^{n} f(Y_{i,t}|Y_{j,t} : (j,t) \in N_{i,t}). \tag{8}$$

Maximization of (8) 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 Splus.

However, the standard errors of these estimates from the standard logistic regression are invalid and hence, need to be recomputed. We use a parametric bootstrap in a similar manner as Gumpertz et al. (1997). In particular, we generate M spatial-temporal binary data sets according to the autologistic model defined in (5), for which the model parameters are fixed at the MPLE $\hat{\boldsymbol{\theta}}$ from the original data. For the m-th data set, 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 error of $\hat{\boldsymbol{\theta}}$.

To generate a spatial-temporal binary data set $\{\boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T\}$ given $\boldsymbol{Y}_1$, we use a Gibbs sampler. By (5), the joint distribution of $\{\boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T\}$ conditional on $\boldsymbol{Y}_1$ is:

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

The normalizing constant $c(\boldsymbol{\theta})$ does not have closed form and hence direct sampling of $\boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T$ from $f(\boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T | \boldsymbol{Y}_1)$ is not possible. Therefore we use Markov chain Monte Carlo, or more specifically, a Gibbs sampler. The main idea is to simulate individual $Y_{i,t}$ from the distribution conditional on all other response variables, for $i = 1, \ldots, n, t = 2, \ldots, T$. Derivation similar as for (6) shows that the full conditional distributions are:

$$\Pr(Y_{i,t} = 1 | \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T\} \setminus Y_{i,t}) = p_{i,t},$$

where for $i = 1, \ldots, n$, the log-odds of success is

$$\log\left(\frac{p_{i,t}}{1 - p_{i,t}}\right) = \begin{cases} \displaystyle\sum_{k=1}^{p} \theta_k X_{k,i,t} + \sum_{j \sim i} \theta_{p+1}(2Y_{j,t} - 1) + \theta_{p+2}(2Y_{i,t-1} + 2Y_{i,t+1} - 2); & 2 \le t < T, \\ \displaystyle\sum_{k=1}^{p} \theta_k X_{k,i,T} + \sum_{j \sim i} \theta_{p+1}(2Y_{j,T} - 1) + \theta_{p+2}(2Y_{i,T-1} - 1); & t = T. \end{cases} \tag{9}$$

It is then straightforward to generate $\{Y_{i,t}\}$ from the full conditional distributions iteratively and after burn-in, we take M samples from the Markov chain as the bootstrap samples. We use the log-posterior values to determine the length of burn-in iterations as in Geweke (1992).

2.2 Optimal Prediction by Markov Chain Monte Carlo

The model defined in (3) and (4) is a vector Markov chain and can provide prediction of $\boldsymbol{Y}_{T+1}, \ldots, \boldsymbol{Y}_{T^*}$ at future time points, given the observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$. By (5), the joint predictive distribution of $\boldsymbol{Y}_{T+1}, \ldots, \boldsymbol{Y}_{T^*}$ conditioned on the observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$ is given by

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

Again we use a Gibbs sampler to draw samples $\boldsymbol{Y}_{T+1}, \ldots, \boldsymbol{Y}_{T^*}$ from the predictive distribution $f(\boldsymbol{Y}_{T+1}, \ldots, \boldsymbol{Y}_{T^*} | \boldsymbol{Y}_T)$. Similar derivation as for (9) gives the full conditional distributions:

$$\Pr(Y_{i,t} = 1 | \{\boldsymbol{Y}_T, \ldots, \boldsymbol{Y}_{T^*}\} \setminus Y_{i,t}) = p_{i,t},$$

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

$$\log\left(\frac{p_{i,t}}{1-p_{i,t}}\right) = \begin{cases} \sum_{k=1}^{p} \theta_k X_{k,i,t} + \sum_{j \sim i} \theta_{p+1}(2Y_{j,t} - 1) + \theta_{p+2}(2Y_{i,t-1} + 2Y_{i,t+1} - 2); & T+1 \le t < T^*, \\ \sum_{k=1}^{p} \theta_k X_{k,i,T^*} + \sum_{j \sim i} \theta_{p+1}(2Y_{j,T^*} - 1) + \theta_{p+2}(2Y_{i,T^*-1} - 1); & t = T^*. \end{cases}$$

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

3 Example: Outbreaks of Southern Pine Beetles

In this section, we apply the spatial-temporal autologistic 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 at a given site to a binomial type of data 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 covariance 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 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 Beetles 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 i-th county and the t-th year. We used the first 31 years (1960–1990) of data for model building and set aside the last 6 years (1991–1996) of data for model validation. Two counties were considered to be neighbors if the corresponding county seats are within 30 miles of each other, as specified 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 demonstrated positive correlations among neighboring counties. Indeed Moran's I index was 0.64 with a p-value less than 0.001 and Geary's c index was 0.32 with a p-value less than 0.001, both indicating strong evidence of positive spatial correlations. 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 to 1996. The epidemic seems to have peaked in the mid-70's and there was evidence of positive correlations 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 m^3/ha), hydric proportion, xeric proportion, size of national forest (in 1000 ha), average daily maximum temperature in the fall (in ^{o}C), average precipitation in the fall (in cm), average daily maximum temperature in the winter (in ^{o}C), average daily maximum temperature in the summer (in ^{o}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. Along with the spatial component and the temporal component, there are a total of 15 variables in the autologistic model (Table 1).

By maximum pseudo-likelihood, the autologistic model had conditional probability:

$$f\left(Y_{i,t} | Y_{j,t} : (j,t) \in N_{i,t}\right) = \frac{\exp\left\{\sum_{k=0}^{13} \hat{\theta}_k X_{k,i,t} Y_{i,t} + \hat{\theta}_{14} \sum_{j \sim i} Y_{i,t}(2Y_{j,t} - 1) + \hat{\theta}_{15} Y_{i,t}(2Y_{i,t-1} - 1)\right\}}{1 + \exp\left\{\sum_{k=0}^{13} \hat{\theta}_k X_{k,i,t} + \hat{\theta}_{14} \sum_{j \sim i}(2Y_{j,t} - 1) + \hat{\theta}_{15}(2Y_{i,t-1} - 1)\right\}}, \tag{10}$$

where $i = 1, \ldots, 100$ and $t = 1960, \ldots, 1990$. Evaluated at the MPLE $\hat{\boldsymbol{\theta}}$, a Gibbs sampler was implemented and after burn-in, a less dependent bootstrap sample of size $M = 1000$ was obtained

by taking every 5-th iteration from the Gibbs sampler. From these 1000 samples, the standard errors of the MPLE $\hat{\boldsymbol{\theta}}$ were estimated. The MPLEs and their corresponding standard errors are reported in Table 1.

Table 1 here

Since not all the coefficients $\{\theta_k : k = 0, \ldots, 15\}$ were significantly different from zero, we set out to determine a suitable reduced model. We started with the full model (10) and performed backward elimination based on a t-ratio between an estimate $\hat{\theta}_k$ and its standard error. At each step, we eliminated the variable that has 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 (see, e.g., Section 11.9, Chatterjee, Hadi, and Price 2000). 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: elevation (X_1) first, then mean summer precipitation (X_{11}), size of national forest (X_6), mean daily maximum fall temperature (X_7), longitude (X_2), interaction between saw volume and mean daily maximum winter temperature (X_{12}), saw volume (X_3), interaction between saw volume and mean daily maximum summer temperature (X_{13}), xeric proportion (X_5), and finally mean daily maximum winter temperature (X_9). The final reduced model has only three explanatory variables, namely the hydric proportion (X_4), the fall precipitation (X_8), and the daily maximum summer temperature (X_{10}). Interestingly both the spatial component and the temporal component are retained. In fact, the spatial-temporal components were the most significant variables throughout the model selection steps. The log-odds of success in the fitted final model is:

$$\begin{aligned}
&\log\left(\frac{\Pr(Y_{i,t}=1|Y_{j,t}:(j,t)\in N_{i,t})}{\Pr(Y_{i,t}=0|Y_{j,t}:(j,t)\in N_{i,t})}\right) \\
&\quad = -17.228 - 0.157\times\sqrt{\text{hydric proportion}} + 0.527\times\text{fall precip} \\
&\qquad + 0.181\times\text{summer temp} + 0.874\times\sum_{j\sim i}(2Y_{j,t}-1) + 0.766\times(2Y_{i,t-1}-1)
\end{aligned} \tag{11}$$

The MPLEs and their corresponding standard errors for the final model are reported in Table 1.

Table 2 here

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. Using the actual data from 1991–1996, we computed the prediction error rates. The results were 0.01, 0.09, 0.06, 0.18, 0.13, and 0.18 for the 6 years. Note that the misclassification rates were larger for the latter three years, possibly because of some larger-than-usual fluctuations in 1994–1996. Nonetheless compared with the results in Table 3 of Gumpertz et al. (2000), our model predicted equally well or better.

4 Discussions

In this article, we have developed an autologistic regression model for binary data which accounts for both spatial dependence and temporal dependence. We have used maximum pseudo-likelihood for parameter estimation and a parametric bootstrap for the corresponding standard errors. Further we have proposed a Gibbs sampler to predict the responses at future time points. The methodology has been applied to successfully predict the outbreaks of Southern Pine Beetles in North Carolina, based on 31 years of data.

Our approach can be extended to Gibbs fields to form more general spatial-temporal auto-models with pairwise spatial dependence. More specifically, suppose a Markov random field in a Gibbsian form:

$$f(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1}) \propto \exp(q(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1}))$$

where

$$q(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1}) = \sum_{i=1}^{n} G_i(Y_{i,t}, \{X_{k,i,t}\}) + \sum_{j\sim i} G_{ij}(Y_{i,t}, Y_{j,t}) + \sum_{i=1}^{n} G_t^*(Y_{i,t}, Y_{i,t-1}).$$

Here $G_i(Y_{i,t}, \{X_{k,i,t}\})$ may correspond to a linear regression, $G_{ij}(Y_{i,t}, Y_{j,t})$ features interactions between the i-th site and j-th site, and the additional $G_t^*(Y_{i,t}, Y_{i,t-1})$ features an interaction between time points t and $t-1$ at the same i-th site. Examples of Gibbs field include auto-binomial, 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-variant coefficients. This can be accomplished by specifying

$$q(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1}) = \sum_{i=1}^{n} G_{it}(Y_{i,t}, \{X_{k,i,t}\}) + \sum_{j\sim i} G_{ijt}(Y_{i,t}, Y_{j,t}) + \sum_{i=1}^{n} G_t^*(Y_{i,t}, Y_{i,t-1}),$$

which we leave for future investigation.

Appendix

In this section, we derive the full conditional distribution in (6) from the joint distribution in (5). Rewrite the full conditional distribution in (6) in terms of the joint distribution in (5):

$$\begin{aligned} & f\left(Y_{i,t}|\{\boldsymbol{Y}_{t_0},\ldots,\boldsymbol{Y}_t\}\setminus Y_{i,t}\right) \\ &= \frac{f\left(\boldsymbol{Y}_{t_0+1},\ldots,\boldsymbol{Y}_t|\boldsymbol{Y}_{t_0}\right)}{f\left(Y_{i,t}=0,\{\boldsymbol{Y}_{t_1},\ldots,\boldsymbol{Y}_t\}\setminus Y_{i,t}|\boldsymbol{Y}_{t_0}\right)+f\left(Y_{i,t}=1,\{\boldsymbol{Y}_{t_1},\ldots,\boldsymbol{Y}_t\}\setminus Y_{i,t}|\boldsymbol{Y}_{t_0}\right)} \\ &\equiv \frac{\exp(A+B(Y_{i,t}))}{\exp(A+B(1))+\exp(A+B(1))} \end{aligned} \tag{12}$$

where

$$\begin{aligned} A \equiv \sum_{t'=t_0+1}^{t}\Bigg(\sum_{i'\neq i}\sum_{k=1}^{p}\theta_k X_{k,i',t'}Y_{i',t'} &+ \sum_{j\sim i'}\theta_{p+1}[Y_{i',t'}Y_{j,t'}+(1-Y_{i',t'})(1-Y_{j,t'})] \\ &+ \sum_{i'\neq i}\theta_{p+2}[Y_{i',t'}Y_{i',t'-1}+(1-Y_{i',t'})(1-Y_{i',t'-1})]\Bigg) \end{aligned}$$

consists of terms in the exponent of (5) not involving $Y_{i,t}$, whereas

$$\begin{aligned} B(Y_{i,t}) \equiv \sum_{k=1}^{p}\theta_k X_{k,i,t}Y_{i,t} &+ \sum_{j\sim i}\theta_{p+1}[Y_{i,t}Y_{j,t}+(1-Y_{i,t})(1-Y_{j,t})] \\ &+\theta_{p+2}[Y_{i,t}Y_{i,t-1}+(1-Y_{i,t})(1-Y_{i,t-1})] \end{aligned}$$

consists of terms in the exponent of (5) involving $Y_{i,t}$. Then,

$$B(0)=\sum_{j\sim i}\theta_{p+1}(1-Y_{j,t})+\theta_{p+2}(1-Y_{i,t-1}),$$

and

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

Hence (12) becomes,

$$\begin{aligned} & \frac{\exp(A+B(Y_{i,t}))}{\exp(A+B(0))+\exp(A+B(1))} \\ &= \frac{\exp(B(Y_{i,t})-B(0))}{1+\exp(B(1)-B(0))} \\ &= \frac{\exp\left\{\sum_{k=1}^{p}\theta_k X_{k,i,t}Y_{i,t}+\sum_{j\sim i}\theta_{p+1}Y_{i,t}(2Y_{j,t}-1)+\theta_{p+2}Y_{i,t}(2Y_{i,t-1}-1)\right\}}{1+\exp\left\{\sum_{k=1}^{p}\theta_k X_{k,i,t}+\sum_{j\sim i}\theta_{p+1}(2Y_{j,t}-1)+\theta_{p+2}(2Y_{i,t-1}-1)\right\}}, \end{aligned}$$

and the full conditional distribution in (6) is obtained.

References

Bailey, R. (1995). Description of the ecoregions of the United States. *Misc. Publication 1391, USDA Forest Service, Washington D.C.*, pp.108.

Besag, J. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society*, Series B, **34**, 75–83.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, Series B, **36**, 192–225.

Besag, J.E. (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*, Third Edition. Wiley, New York.

Cressie, N. (1993). *Statistics for Spatial Data*, Revised Edition. Wiley, New York.

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 University Press, Oxford.

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*, Series 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.

Huffer, F. 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. *Res. Population Ecology*, **26**, 261–274.

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.

Table 1: Maximum pseudo-likelihood estimates of the coefficients for the full model and the final model selected by backward elimination. Values in the parentheses are the standard errors estimated by a bootstrap method.

Variable	Full Model		Final Model	
X_0 Intercept	-33.237	(29.947)	-17.228	(10.431)
X_1 Ln[elevation (m)]	0.007	(0.042)	—	—
X_2 Longitude	0.111	(0.259)	—	—
X_3 $\sqrt{\text{saw volume (m}^3\text{/ha)}}$	2.466	(2.679)	—	—
X_4 $\sqrt{\text{hydric proportion}}$	-0.116	(0.113)	-0.157	(0.047)
X_5 $\sqrt{\text{xeric proportion}}$	0.064	(0.074)	—	—
X_6 $\sqrt{\text{national forest (thousand ha)}}$	0.032	(0.087)	—	—
X_7 Mean daily maximum fall temp (C)	-0.348	(0.750)	—	—
X_8 Mean fall precipitation (cm)	0.690	(0.574)	0.527	(0.417)
X_9 Mean daily maximum winter temp (C)	-0.230	(0.453)	—	—
X_{10} Mean daily maximum summer temp (C)	0.888	(0.689)	0.181	(0.105)
X_{11} Mean summer precipitation (cm)	0.055	(0.417)	—	—
$X_{12} = X_3 \times X_9$	0.025	(0.031)	—	—
$X_{13} = X_3 \times X_{10}$	-0.044	(0.040)	—	—
Spatial effect	0.883	(0.159)	0.874	(0.148)
Temporal effect	0.751	(0.235)	0.766	(0.228)

Table 2: Individual steps in a backward elimination. Reported are the maximum pseudo-likelihood estimates of the coefficients and the standard errors (in parentheses) estimated by a bootstrap method.

Variable	Step 0		Step 1		Step 2		Step 3	
X_0	−33.237	(29.947)	−33.515	(28.841)	−32.277	(26.186)	−32.692	(26.082)
X_1	0.007	(0.042)	—	—	—	—	—	—
X_2	0.111	(0.259)	0.106	(0.194)	0.113	(0.190)	0.080	(0.167)
X_3	2.466	(2.679)	2.453	(2.523)	2.331	(2.296)	2.191	(2.175))
X_4	−0.116	(0.113)	−0.116	(0.116)	−0.115	(0.114)	−0.114	(0.111)
X_5	0.064	(0.074)	0.064	(0.074)	0.066	0.073)	0.069	(0.070)
X_6	0.032	(0.087)	0.032	(0.084)	0.035	(0.076)	—	—
X_7	−0.348	(0.750)	−0.351	(0.743)	−0.370	(0.712)	−0.252	(0.633)
X_8	0.690	(0.574)	0.691	(0.598)	0.709	(0.576)	0.694	(0.564)
X_9	−0.230	(0.453)	−0.230	(0.459)	−0.204	(0.396)	−0.227	(0.385)
X_{10}	0.888	(0.689)	0.890	(0.694)	0.885	(0.676)	0.775	(0.580)
X_{11}	0.055	(0.417)	0.054	(0.408)	—	—	—	—
X_{12}	0.025	(0.031)	0.025	(0.031)	0.024	(0.030)	0.022	(0.029)
X_{13}	−0.044	(0.040)	−0.044	(0.039)	−0.042	(0.036)	−0.039	(0.033)
X_{14}	0.883	(0.159)	0.883	(0.162)	0.883	(0.163)	0.882	(0.162)
X_{15}	0.751	(0.235)	0.751	(0.239)	0.752	(0.237)	0.751	(0.228)

Table 2: Individual steps in a backward elimination. Reported are the maximum pseudo-likelihood estimates of the coefficients and the standard errors (in parentheses) estimated by a bootstrap method.

Variable	Step 4		Step 5		Step 6		Step 7	
X_0	−32.273	(25.676)	−36.386	(19.543)	−32.081	(18.675)	−24.657	(13.551)
X_1	—	—	—	—	—	—	—	—
X_2	0.059	(0.158)	—	—	—	—	—	—
X_3	2.022	(2.086)	1.962	(1.999)	1.558	(1.963)	—	—
X_4	−0.121	(0.109)	−0.096	(0.082)	−0.089	(0.081)	−0.095	(0.079)
X_5	0.066	(0.069)	0.048	(0.050)	0.042	(0.049)	−0.039	(0.048)
X_6	—	—	—	—	—	—	—	—
X_7	—	—	—	—	—	—	—	—
X_8	0.649	(0.542)	0.639	(0.535)	0.651	(0.504)	−0.710	(0.490)
X_9	−0.332	(0.312)	−0.298	(0.294)	−0.145	(0.127)	−0.137	(0.126)
X_{10}	0.605	(0.348)	0.579	(0.336)	0.435	(0.249)	−0.341	(0.196)
X_{11}	—	—	—	—	—	—	—	—
X_{12}	0.020	(0.029)	0.017	(0.027)	—	—	—	—
X_{13}	−0.036	(0.031)	−0.033	(0.030)	−0.018	(0.023)	−0.000	(0.001)
X_{14}	0.880	(0.158)	0.879	(0.161)	0.878	(0.154)	0.880	(0.151)
X_{15}	0.749	(0.218)	0.745	(0.227)	0.749	(0.227)	0.763	(0.291)

Table 2: Individual steps in a backward elimination. Reported are the maximum pseudo-likelihood estimates of the coefficients and the standard errors (in parentheses) estimated by a bootstrap method.

Variable	Step 8		Step 9		Step 10	
X_0	-24.048	(13.099)	-20.943	(12.447)	-17.228	(10.431)
X_1	—	—	—	—	—	—
X_2	—	—	—	—	—	—
X_3	—	—	—	—	—	—
X_4	-0.102	(0.067)	-0.117	(0.063)	-0.157	(0.047)
X_5	0.038	(0.047)	—	—	—	—
X_6	—	—	—	—	—	—
X_7	—	—	—	—	—	—
X_8	0.711	(0.483)	0.672	(0.483)	0.527	(0.417)
X_9	-0.132	(0.121)	-0.102	(0.112)	—	—
X_{10}	0.329	(0.184)	0.278	(0.170)	0.181	(0.105)
X_{11}	—	—	—	—	—	—
X_{12}	—	—	—	—	—	—
X_{13}	—	—	—	—	—	—
X_{14}	0.881	(0.150)	0.874	(0.154)	0.874	(0.148)
X_{15}	0.763	(0.238)	0.761	(0.262)	0.766	(0.228)

Figure 1: Total number of years a county experienced an outbreak of Southern Pine Beetles in 1960–1996, for each of the 100 counties of North Carolina.

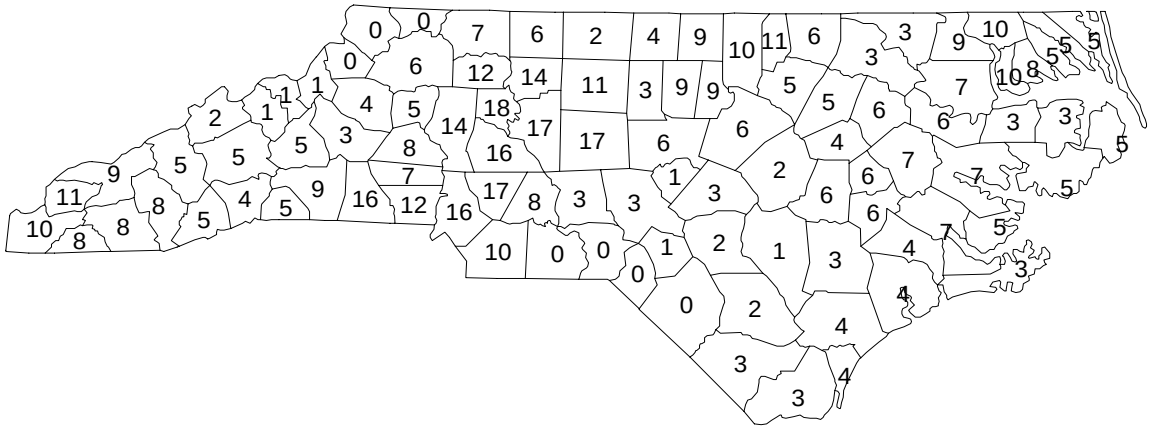

Figure 2: Total number of counties that experienced an outbreak of Southern Pine Beetles in the state of North Carolina from 1960 to 1996.